UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Plataforma de Indicadores Académicos

Rebeca Maria Cantos de Atouguia

Mestrado em Gestão de Informação
Área de Especialização em Gestão e Análise de Dados

Trabalho de Projeto orientado por:
Professor Doutor António Manuel Silva Ferreira
Professora Doutora Maria Fernanda Oliveira

2018

Agradecimentos

Gostaria de agradecer a todas as pessoas que contribuíram para a realização deste trabalho.

Aos meus orientadores, Professor António Ferreira e Professora Fernanda Oliveira, pela excelente orientação que me proporcionaram e pela total disponibilidade desde o primeiro momento.

Ao Professor António Ferreira, por me ter transmitido conhecimentos de forma tão exemplar e discreta, numa área que não me era familiar, e por todos os contributos na escrita do relatório.

À Professora Fernanda Oliveira, pela motivação e confiança transmitidas, pelos valiosos ensinamentos na aplicabilidade deste projeto, e por todo o apoio manifestado para a concretização deste trabalho.

Aos professores da FCUL, pela formação de excelência que me proporcionaram ao longo do meu percurso académico.

A todos os docentes e colegas da FCUL, por contribuírem para o meu crescimento profissional.

Aos meus verdadeiros amigos, por estarem sempre presentes.

À minha família, pelo carinho e proximidade, apesar da distância.

Aos meus Pais, à minha irmã e cunhado, por serem um exemplo de vida.

Resumo

Cada vez é mais importante que as organizações disponham de informação de qualidade que lhes permita obter vantagens competitivas e alcançar os objetivos propostos. Não é suficiente ter uma grande quantidade de dados, que crescem continuamente, armazenados nos sistemas operacionais. É necessário compreender o seu significado e transformá-los em informação, de modo a conseguir responder às perguntas dos decisores e apoiar a tomada de decisões. Também é importante obter a informação relevante e precisa em tempo útil e disponibilizá-la de uma forma facilmente compreensível. Adicionalmente, se precisamos dessa informação periodicamente, é conveniente automatizar e melhorar processos para permitir que os recursos humanos se dediquem à análise da informação e não tanto à sua geração.

O propósito deste projeto foi disponibilizar, em ambiente *web*, uma plataforma de indicadores académicos da Faculdade de Ciências da Universidade de Lisboa (FCUL). O trabalho teve uma série de etapas, iniciando-se com a modelação dimensional dos dados, seguindo-se a extração de dados de diferentes fontes, a sua conformação e carregamento num repositório único, e finalizando com a elaboração de relatórios. O grande volume de dados que é gerado num estabelecimento de ensino superior como a FCUL, torna cada vez mais útil e necessária a existência de uma plataforma de visualização de informação, que possa dar resposta quer a pedidos de reporte institucional e de apoio à decisão, quer à comunidade de Ciências em geral.

O primeiro objetivo do trabalho foi pesquisar conceitos sobre *Business Intelligence* (BI), modelação dimensional de dados, técnicas de visualização de informação e, por último, a escolha da ferramenta para suportar o processo de transformação de dados e produção de relatórios, até serem disponibilizados os indicadores académicos na *web*.

O segundo objetivo consistiu na obtenção de dados de diferentes fontes, e sua conformação e transformação, para posterior carregamento num repositório único. A sequência de transformações utilizada para converter dados em informação útil ficou automatizada de modo a poder ser repetida sempre que forem carregados dados novos, permitindo que a informação seja disponibilizada de forma mais célere. Este objetivo foi avaliado através da obtenção, transformação e carregamento de dados referentes ao ano letivo 2017/18 e verificando a utilidade das rotinas automatizadas e a redução do tempo despendido.

Por último, no terceiro objetivo, foram elaborados relatórios com indicadores académicos, possibilitando não só a obtenção de dados sobre um ano letivo, em particular o que estiver em curso, mas também a análise de tendências. A disponibilização em ambiente *web* dessa informação à comunidade de Ciências, permite a cada utilizador (órgãos de gestão, presidentes de departamento, coordenadores de cursos, docentes e funcionários não docentes) a consulta de acordo com o seu interesse particular, com maior ou menor nível de detalhe e de uma forma interativa. A avaliação destes relatórios foi feita, quer através da validação das respostas às perguntas de negócio identificadas na definição de requisitos, quer pela análise de dois decisores, confirmando-se que satisfaz as suas necessidades.

A realização deste trabalho foi uma oportunidade de desenvolver várias competências e adquirir conhecimentos essenciais para aplicar no futuro, nomeadamente em contexto laboral.

Palavras-Chave: *Business Intelligence*, Modelação dimensional, Indicadores académicos, Relatórios interativos

Abstract

It is becoming more and more important for organizations to have quality information that allows them to gain competitive advantages and achieve the proposed objectives. It is not enough to have a lot of data, which grows continuously, stored in operational systems. It is necessary to understand the meaning of the data and turn them into information to be able to answer the questions of decision-makers. It is also important to get relevant and accurate information in a timely manner and to make it available in an easily understandable way. In addition, if we need that information, from time to time, it is convenient to automate and improve processes, to allow human resources to focus on the information itself rather than its generation.

The purpose of this project was to provide the Faculty of Sciences of the University of Lisbon (FCUL) with a platform of academic indicators, in a web environment. The work comprised a number of stages, starting with the dimensional modeling of the data, followed by the extraction of data from different sources, its conformation and loading into a single repository, and ending with the elaboration of reports. The large amount of data that is generated in a higher education institution, such as FCUL, makes it increasingly useful and necessary to have an information visualization platform, whether to enable institutional reporting and decision making, or to just serve the academic community of "Ciências", in general.

The first goal of the work was to research concepts about Business Intelligence (BI), dimensional modeling, information visualization techniques and, lastly, the choice of tool to support the process of data transformation and production of reports on the web.

The second goal consisted in obtaining data from different sources, and their conformation and transformation, for subsequent loading into a single repository. The transformation sequence used to convert data into useful information was automated so it can be repeated whenever new data is loaded, allowing the information to become available more quickly. This objective was evaluated by obtaining, transforming and loading data for the 2017/18 school year and by verifying the usefulness of the automated routines and the reduced work timespan.

Finally, for the third goal, reports were prepared with academic indicators, making it possible not only to obtain data about a school year, specifically the current one, but also the analysis of trends. The availability of this information in a web environment to the "Ciências" community allows each user (management bodies, department presidents, course coordinators, teachers and non-teaching staff) to consult according to their particular interests, with a higher or lower level of detail and in an interactive manner. The evaluation of these reports took place by validating the answers to the business questions identified in the requirements definition and through the opinions of two decision-makers, confirming that the developed platform met their needs.

This undertaking of this project was an opportunity to develop various skills and acquire essential knowledge to apply in the future, especially in a work context.

Keywords: Business Intelligence, Dimensional modeling, Academic indicators, Interactive reports

Índice

Lista de figuras

Lista de tabelas

Lista de acrónimos

Acrónimo	Significado
AEPQ	Área de Estudos, Planeamento e Qualidade
A3ES	Agência de Avaliação e Acreditação do Ensino Superior
BI	*Business Intelligence*
CNA	Concurso Nacional de Acesso
CNAEF	Classificação Nacional de Áreas de Educação e Formação
DAX	*Data Analysis eXpressions*
DGEEC	Direção-Geral das Estatísticas de Educação e Ciência
DGES	Direção-Geral do Ensino Superior
ECTS	*European Credit Transfer System*
ETL	*Extract, Transform and Load*
EUROSTAT	Autoridade Estatística da União Europeia
FCUL	Faculdade de Ciências da Universidade de Lisboa
GAAI	Gabinete de Avaliação e Auditoria Interna
GOGI	Gabinete de Organização e Gestão de Informação
IES	Instituição de Ensino Superior
NUTS	Nomenclatura das Unidades Territoriais para Fins Estatísticos
OCDE	Organização para a Cooperação e Desenvolvimento Económico
PRIES	Plataforma de Recolha de Informação do Ensino Superior
RAIDES	Registo de Alunos Inscritos e Diplomados no Ensino Superior
SAD	Sistema de Apoio à Decisão
SQL	*Structured Query Language*
TI	Tecnologias de Informação
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
XML	*eXtensible Markup Language*

1. Introdução

Para Andreas Schleicher, responsável pela área da Educação na OCDE (Organização para a Cooperação e Desenvolvimento Económico), numa entrevista ao Semanário Expresso, em abril de 2016, "O mundo já não recompensa as pessoas apenas por aquilo que sabem - o Google sabe tudo - mas por aquilo que conseguem fazer com isso. Por isso a educação tem cada vez mais que ver com o desenvolvimento da criatividade, do pensamento crítico, da resolução de problemas e da tomada de decisões." [1]

Um Sistema de Apoio à Decisão (SAD), também designado por sistema de *Business Intelligence* (BI), compreende diversas técnicas que permitem que uma organização disponha de informação útil para a tomada de decisões. Foi com este tipo de sistema que decorreu o trabalho descrito neste relatório, no âmbito do Projeto em Gestão de Informação. No presente capítulo descreve-se a motivação para o projeto, apresentam-se os objetivos definidos, faz-se uma descrição da instituição, apresentam-se as contribuições do trabalho realizado, a notação adotada e a estrutura do documento.

1.1 Motivação

Um dos ativos mais importantes que uma organização possui são os seus dados. Devido ao contínuo desenvolvimento dos sistemas informáticos nas organizações, tem havido um grande crescimento no que diz respeito à geração e armazenamento de dados provenientes dos sistemas transacionais, isto é, que suportam o funcionamento operacional da organização. Assim, as bases de dados têm crescido a um ritmo que muitas vezes excede a capacidade de interpretar e compreender tanta informação [2].

Isto também acontece em instituições de ensino superior (IES), onde os sistemas transacionais e as bases de dados são parte da gestão administrativa, uma vez que as unidades académicas não poderiam funcionar sem aplicações que registem a grande quantidade de alunos inscritos, os seus cursos, as notas nas unidades curriculares, bem como um extenso número de variáveis que têm de ser produzidas e analisadas para cada um deles (por exemplo, o número de inscrições de um aluno).

Tendo em conta a dimensão da Faculdade de Ciências da Universidade de Lisboa (FCUL), com mais de 5000 alunos inscritos e cerca de 80 cursos conferentes de grau, o desafio abordado neste trabalho é o de tirar o melhor partido possível dessa grande quantidade de dados para melhorar a qualidade dos processos e serviços que a instituição deve oferecer. Por outro lado, é fundamental facilitar o acesso a essa informação a todas as partes interessadas, de forma correta e atempada. Disponibilizando a referida informação é mais fácil tomar decisões e responder às solicitações legais impostas pela tutela ou por outros organismos oficiais.

De acordo com a Lei nº 38/2007 de 16 de agosto, que aprova o regime jurídico de avaliação do ensino superior [3], um dos objetivos da avaliação da qualidade é a prestação de informação fundamentada à sociedade sobre o desempenho das IES. De acordo com a mesma lei, no âmbito da respetiva autoavaliação, as IES devem: i) "certificar-se de que recolhem, analisam e usam a informação relevante para a gestão eficaz dos seus ciclos de estudos e de outras atividades" e ii) "publicar, regularmente, informação quantitativa e qualitativa, atualizada, imparcial e objetiva acerca dos ciclos de estudos que ministram e graus e diplomas que conferem".

Utilizando as ferramentas apropriadas de análise, é mais fácil gerir os dados, transformando-os em informação e essa informação em conhecimento. Para conseguir esse objetivo é fundamental a utilização de tecnologias adequadas, nomeadamente de BI.

Com um sistema de BI, pretendeu-se com este projeto divulgar um conjunto de indicadores, em ambiente *web*, sobre cursos e alunos da FCUL, obtidos de diferentes fontes de dados. Esta disponibilização da informação deve poder ser feita através de relatórios interativos, que permitam aos vários perfis de utilizadores (órgãos de gestão, docentes, investigadores, funcionários não docentes e alunos) analisá-la de diferentes perspetivas, podendo personalizar as visualizações com maior ou menor detalhe.

1.2 Objetivos

Os principais objetivos deste projeto foram os seguintes:

Objetivo 1 (O1) - Conceitos e familiarização com as tecnologias: pesquisa de conceitos sobre BI, fundamentos da modelação de dados, das técnicas de visualização da informação e escolha da ferramenta de *BI & Analytics* para este projeto. Conhecendo os princípios fundamentais de um sistema de BI, conseguem-se melhores resultados, independentemente da ferramenta a utilizar.

Objetivo 2 (O2) - Conformação dos dados e automação de consultas: a obtenção de dados de diferentes fontes implica uma transformação dos mesmos, para posterior carregamento num repositório único. Para isso, devem ser estudadas diversas técnicas de forma a minimizar problemas de incoerência dos dados: a uniformização de diferentes fontes, o tratamento de duplicados, a correção de erros, o preenchimento de valores em falta, a uniformização de nomes e unidades de medida. Esta sequência de transformações para converter dados em informação útil, deve ficar automatizada de modo a poder ser repetida sempre que são carregados dados novos. Este objetivo foi avaliado comparando este novo processo com o anterior, através da obtenção, transformação e carregamento de novos dados referentes ao ano letivo 2017/18 e verificando a utilidade das rotinas automatizadas.

Objetivo 3 (O3) - Produção e publicação de relatórios: A visualização da informação, em relatórios interativos, deve permitir aos órgãos de gestão, presidentes de departamento, coordenadores de cursos, docentes e funcionários não docentes de departamentos e unidades serviço, fazer uma análise exploratória dos dados em diferentes perspetivas. A avaliação destes relatórios teve duas componentes: a forma como a informação foi apresentada, através da verificação do cumprimento das recomendações de elaboração de relatórios, e o conteúdo da informação, através da validação das respostas às perguntas de negócio. Existiu ainda uma avaliação por parte de dois decisores que confirmaram a sua pertinência.

Foram ainda definidos dois pré-requisitos para a realização deste trabalho: a ferramenta de BI escolhida não implicar nenhum custo financeiro para a instituição, e a autonomia funcional em relação à equipa de Tecnologias de Informação (TI) da FCUL, relativamente à obtenção dos dados, ao processo de ETL (*Extract, Transform and Load*), ao repositório dos dados e à geração de relatórios.

1.3 Descrição da instituição

Contabilizando discentes, docentes, investigadores e funcionários não docentes, a FCUL ultrapassa as 6000 pessoas. Dispõe de um conjunto de 12 unidades de serviço (estruturas de apoio logístico, técnico e administrativo), que permitem o desempenho das funções e dos objetivos a que a Faculdade se propõe. A Área de Estudos, Planeamento e Qualidade (AEPQ) é constituída pelo Gabinete de Organização e Gestão de Informação (GOGI) e pelo Gabinete de Avaliação e Auditoria Interna (GAAI), no qual estou inserida. Este gabinete tem como principais atribuições: o tratamento de dados estatísticos e de inquéritos de satisfação, a monitorização dos principais indicadores da FCUL, que se encontram em fase de uniformização no Manual da Qualidade de Ciências, a elaboração de estudos sobre o sucesso escolar e o concurso nacional de acesso, e o acompanhamento dos processos de acreditação/avaliação de ciclos de estudos e do sistema integrado de garantia da qualidade.

1.4 Contribuições

A concretização do objetivo **O1 - Conceitos e familiarização com as tecnologias** permitiu adquirir conhecimentos de BI, fundamentais para quem trabalha com grandes quantidades de dados. Permitiu ainda a obtenção dos fundamentos da modelação de dados, das técnicas de visualização da informação e a familiarização com a ferramenta escolhida. Esta familiarização possibilitou mostrar a ferramenta e uma utilização prática em contexto de trabalho, numa aula de Integração e Processamento Analítico de Informação, em maio de 2017.

O objetivo **O2 - Conformação dos dados e automação de consultas** permitiu corrigir situações de incoerência dos dados, nomeadamente no que diz respeito aos dados sobre os alunos, uma vez que cobrem vários anos letivos. Todas as transformações realizadas ficaram gravadas e automatizadas, para utilização futura, permitindo que a informação seja disponibilizada de forma mais célere. Este objetivo permitiu reduzir o tempo e o esforço despendidos na realização destas tarefas e minimizar erros que poderiam ocorrer durante a execução manual das mesmas.

A realização do objetivo **O3 - Produção e publicação de relatórios** possibilitou analisar tendências e evoluções através de relatórios interativos, os quais foram avaliados por dois decisores que confirmaram a sua pertinência. A disponibilização em ambiente *web* de indicadores académicos à comunidade de Ciências permitiu que cada utilizador os possa consultar de acordo com o seu interesse particular e de forma interativa. Estes relatórios incluem informação mais concisa, mais específica para as diferentes partes interessadas e melhor apresentada.

1.5 Notação adotada

Este relatório foi escrito em português, ao abrigo do novo acordo ortográfico, e todos os termos de outro idioma apresentam-se em itálico.

1.6 Estrutura do documento

O relatório está estruturado nos seguintes cinco capítulos:

No Capítulo 1 descreve-se a motivação para o projeto, apresentam-se os objetivos definidos, faz-se uma descrição da instituição, apresentam-se as contribuições do projeto e a notação adotada.

No Capítulo 2 expõem-se conceitos teóricos relacionados com o tema do projeto, nomeadamente sobre informação e dados, *business intelligence*, *data warehouse e data marts*, modelação dimensional, visualização da informação e apresenta-se a ferramenta de BI utilizada no projeto.

No Capítulo 3 apresenta-se a análise do problema, com a definição de requisitos gerais, os processos de negócio, a descrição das fontes de dados e dos indicadores académicos.

No Capítulo 4 descreve-se detalhadamente a concretização da solução, nomeadamente a modelação dimensional da plataforma de indicadores académicos; a obtenção dos dados das diferentes fontes, a sua integração e transformação; o enriquecimento do modelo e obtenção de cálculos analíticos e, por último, a visualização da informação através da elaboração de relatórios, incluindo a apresentação de alguns dos relatórios obtidos.

No Capítulo 5 discutem-se as principais contribuições do trabalho realizado, as competências adquiridas no decorrer do projeto, elencam-se as principais dificuldades encontradas, e são apontadas linhas para trabalho futuro.

2. Conceitos

Neste capítulo são abordados conceitos teóricos fundamentais para a realização deste projeto, obtidos por pesquisa da literatura, nomeadamente a importância da informação, o significado de *Business Intelligence* (BI), a distinção entre *data warehouse e data marts* e a modelação dimensional de dados. Também são mencionadas recomendações de apresentação da informação e é documentado o processo de escolha da ferramenta de BI utilizada neste trabalho.

2.1 Informação e dados

A informação é um dos recursos mais importantes de uma organização, contribuindo decisivamente para a sua maior ou menor competitividade. Com o aumento da concorrência, tornou-se vital melhorar as capacidades de decisão a todos os níveis. Atualmente, a tomada de decisão nas organizações é um processo complexo, dada a quantidade de informação em jogo, a sua complexidade e a frequência com que se altera. Para que a informação possa ser utilizada como meio eficaz à tomada de decisão, deve verificar simultaneamente as seguintes condições [2]:

- Ser atual: só com base em informação atualizada se podem tomar decisões acertadas;
- Ser rigorosa: só com informação correta e precisa se pode decidir com confiança;
- Ser relevante: a informação deve ser devidamente filtrada, dado o grande volume de informação envolvida no processo de tomada de decisão;
- Ser disponibilizada em tempo oportuno: a sua utilidade poderá ser posta em causa se não puder estar disponível no momento em que é solicitada;
- Ser inteligível: a informação só é informação se puder ser interpretada e entendida.

Numa organização podem ser identificados dois tipos de informação [4]:

- Operacional: utilizada diariamente, permitindo que organização leve a cabo as suas atividades de rotina de forma eficiente;
- De gestão: serve de suporte à tomada de decisão nos três níveis da gestão (operacional, tático e estratégico).

Um outro conceito relacionado com a informação é o conceito de dados. Dados e informação são coisas distintas: dados são apenas elementos ou valores concretos que isoladamente não têm qualquer valor. Os dados só se transformam em informação quando relacionados ou interpretados de alguma forma [2].

O ciclo dados-informação-decisão [4], representado na Figura 2.1, existe em qualquer organização e pode ser descrito da seguinte forma: o utilizador aplica a sua inteligência sobre os dados produzindo informação. Esta é a base do conhecimento que é utilizado na tomada de decisões. As decisões geram determinadas ações que produzem mais dados. E o ciclo volta ao seu início.

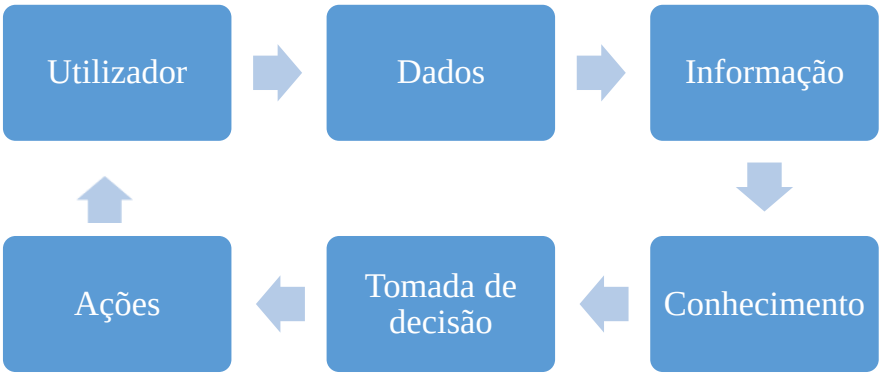


Figura 2.1 - Ciclo dados-informação-decisão [4]

Desta forma**,** as organizações geram diariamente um volume de dados muito elevado, acrescido de dados gerados externamente à organização e que, em muitas situações, também devem ser analisados em conjunto com os dados internos, para obter informação complementar.

2.2 *Business intelligence*

O termo *Business Intelligence* (inteligência de negócio), conhecido habitualmente como BI, é muito utilizado hoje em dia. Existem inúmeras definições para este conceito, sendo neste trabalho usada a seguinte: trata-se de um conjunto de estratégias, tecnologias e metodologias que ajuda a transformar os dados em informação de qualidade, e a informação em conhecimento, permitindo tomar decisões de forma mais acertada e ajudando a melhorar a competitividade [5].

A definição anterior reforça a ideia que BI não é só tecnologia. É também metodologia e, ao nível do negócio, tem que estar alinhado com a estratégia da instituição. Por este motivo uma das primeiras perguntas antes de começar a desenvolver um sistema de BI é: onde queremos chegar?

Um sistema de BI é um sistema de informação de onde se obtém a informação que a organização precisa. Neste sistema ou repositório, os dados procedentes de diferentes aplicações, base de dados, ficheiros, páginas e serviços *web* já não estão no seu formato original, uma vez que sofreram uma série de processos de transformação e limpeza, e têm maior qualidade. Estão preparados para poder dar resposta de uma forma eficaz, rápida e certeira às perguntas de negócio realizadas pelos utilizadores.

Os sistemas de BI, designados nas décadas de 80 e 90 por sistemas de informação para executivos, têm como principais características [6]:

- Serem adaptáveis a necessidades individuais;
- Permitirem a análise e navegação exploratória dos dados;
- Admitirem a aplicação de filtros e agregações hierárquicas;
- Cruzar várias fontes de dados;
- Acompanhar tendências e sinalizar exceções;
- Apresentar a informação de forma gráfica, tabular e textual.

Os sistemas transacionais ou operacionais têm como objetivo fazer os dados entrar na organização e garantir o seu funcionamento, suportando processos curtos e repetitivos de escrita e leitura de dados. Por outro lado, os sistemas analíticos ou de apoio à decisão, onde se enquadram os sistemas de BI, reúnem dados provenientes de várias fontes, e servem para facilitar a navegação e análise dos dados, através de processos longos e exploratórios de leitura de dados.

Assim, os sistemas de BI não são um substituto dos sistemas operacionais já existentes, nem pelo facto de desenvolver um sistema de BI as formas que a organização já tem de obter informação deixam de ter utilidade. São sistemas com objetivos diferentes, mas que se devem complementar para otimizar o valor dos sistemas de informação.

A Tabela 2.1 apresenta as principais diferenças entre dados operacionais e dados de apoio à decisão:

Tabela 2.1 - Caraterísticas dos dados operacionais e de apoio à decisão [7]

Dados operacionais	Dados de apoio à decisão
Orientados a transações	Orientados a análises
Processamento repetitivo	Processamento exploratório
Poucos dados acedidos	Muitos dados acedidos
Nomes curtos e códigos	Nomes inteligíveis
Detalhados	Detalhados e agregados
Sobre o estado atual	Sobre a evolução histórica
Atualizados em contínuo	Atualizações planeadas
Fontes de dados internas	Fontes de dados internas e externas
Relatórios pré-definidos	Relatórios personalizados

Os dados operacionais estão orientados a transações, como, por exemplo, uma inscrição de um aluno numa unidade curricular. Os dados de apoio à decisão estão orientados à análise, para detetar padrões e exceções nos dados, e permitem navegações exploratórias sobre os dados em vários níveis de detalhe. Nos dados operacionais, em que predomina o processo de escrita, são acedidos poucos dados, enquanto nos de apoio à decisão, ao predominar o processo de leitura, o número de dados acedidos é elevado. Neste último tipo de dados, os nomes são inteligíveis para estar ao alcance dos decisores, e têm agregações pré-calculadas, que permitem resposta rápida a interrogações complexas.

Outra diferença tem a ver com o histórico e as fontes de dados: os dados operacionais referem-se à situação atual e são provenientes dos sistemas operacionais internos, enquanto os de apoio à decisão são referentes a longos períodos de tempo, permitindo analisar tendências, e são provenientes de várias fontes, tanto internas como externas. Os dados transacionais são atualizados em contínuo e nos analíticos as alterações são graduais e planeadas, para evitar invalidar dados existentes.

Os relatórios dos dados operacionais são pré-definidos por uma questão de desempenho, com configurações fixas e pouco flexíveis, com informação estática e que nem sempre responde às necessidades. Nos dados de apoio à decisão os relatórios são dinâmicos, interativos, com desenho flexível dos ecrãs, podendo ser personalizados, e são criados de forma rápida e simples, uma vez que o modelo de dados é mais simples e com menos tabelas que o modelo dos sistemas operacionais.

Apesar de existirem muitas variantes nas arquiteturas de um sistema de BI, os principais elementos que o compõem são os seguintes:

- **Fontes de dados**: são as bases de dados dos sistemas operacionais (internos), páginas e serviços *web*, fontes *open data*, ficheiros de texto, folhas de cálculo;
- ***Data warehouse***: é o repositório/ base de dados com informação já trabalhada a partir das fontes de dados do ponto anterior. Para a carregar com informação, executam-se processos de forma

periódica que se encarregam da extração de dados das fontes, transformação em informação e carregamento no repositório, o que se designa de processo de ETL (*Extract, Transform and Load*). O processo de ETL é umas das atividades técnicas mais críticas do desenvolvimento de uma solução de BI e da sua adequada implementação vai depender a coerência dos dados utilizados nas análises de dados.

- **Ferramentas analíticas**: servem para o utilizador poder visualizar e analisar a informação, podendo interagir com ela e utilizá-la como apoio na tomada de decisões.

A Figura 2.2 apresenta a arquitetura de um sistema de BI: os dados extraídos dos sistemas operacionais e diferentes fontes são guardados na *data staging area*, onde sofrem processos de correção de incoerências, formatação e transformação (ETL). Quando estão prontos a usar, são carregados no *data warehouse* para que possam ser acedidos pelos utilizadores no apoio à tomada de decisões analíticas.

Figura 2.2 - Principais elementos da arquitetura de um sistema de BI [Adaptado de 8]

A componente referente ao *data warehouse* é apresentada na secção seguinte, e a componente analítica relativa à apresentação e visualização da informação, na Secção 2.5.

2.3 *Data warehouse* e *data marts*

Um dos elementos principais de um sistema de BI é o *data warehouse* (armazém de dados). Na década de 90, Bill Inmon, considerado o "pai dos *data warehouses*" definiu um *data warehouse* da seguinte forma: uma coleção de dados orientados por assuntos, integrados, não voláteis e variáveis no tempo, organizados para suportar necessidades empresariais [9]. Cada um dos pontos desta definição é detalhado a seguir:

- **Orientados por assuntos**: focados nos aspetos de negócio e excluindo os dados que não são relevantes para o processo de tomada de decisão. Ou seja, a informação é classificada com base nos interesses da organização;
- **Integrados:** os dados provêm de diversas fontes, com os mais diversos formatos, pelo que devem ser consolidados e integrados antes de serem carregados. A conformação resolve problemas relacionados com as convenções de nomes, unidades de medida, codificações, fontes

de diferentes dados, entre outros. O processo de ETL é uma das etapas mais importantes da construção de um *data warehouse* que requer atenção, tempo e qualidade;

- **Não-voláteis**: os dados inseridos não devem ser modificados ou removidos, pois a informação é útil para as análises e a tomada de decisão quando é estável. Os dados operacionais variam constantemente, mas os dados depois de entrar no *data warehouse* já não mudam;
- **Variáveis no tempo**: um *data warehouse* mantém a informação atual e os dados históricos. Este conceito de estrutura temporal permite detetar padrões e relações a longo prazo para auxiliar na tomada de decisão. O tempo é uma dimensão essencial que todos os *data warehouse* devem suportar.

Os *data warehouses* para além da função de consolidação de dados de várias fontes num repositório, servem para dar resposta rápida a interrogações complexas, permitindo realizar análises exploratórias de dados em vários níveis de detalhe e com a possibilidade de aplicar filtros sobre os valores.

Segundo Kimball [10], um *data warehouse* tem como principais objetivos:

- Tornar a informação facilmente acessível: o conteúdo do *data warehouse* deve ser intuitivo e de fácil entendimento para o utilizador que toma as decisões. Devem ser usados termos conhecidos pelos decisores, permitindo navegações exploratórias sobre os mesmos;
- Apresentar informações coerentes: a informação deve ter qualidade e todos os dados devem ser relevantes, precisos e completos. Para que sejam guardados dados coerentes, é necessária uma fase prévia de tratamento de dados duplicados, com erros e desatualizados, bem como a consolidação de múltiplas fontes de dados;
- Ser adaptativo e tolerante à mudança: as alterações no *data warehouse* devem ser suaves, não quebrando a compatibilidade com dados e aplicações existentes, nem invalidando relatórios já construídos;
- Auxiliar no processo de tomada de decisão e ser aceite pela organização: de nada adianta construir uma solução que não é utilizada, por não fornecer os indicadores necessários à tomada de decisão. Um *data warehouse* deve guardar os dados certos para a tomada de decisão.

Os *data marts* são repositórios de informação que incidem sobre um dos processos de negócio, conforme apresentado na Figura 2.3. São mais pequenos, mais focados e mais eficientes. Os vários *data marts* constituem o *data warehouse*.

Um processo de negócio é um conjunto de atividades definidas dentro de uma empresa num tema específico. Usualmente os processos de negócio são suportados por sistemas operacionais e possuem medidas específicas, ligadas ao desempenho organizacional.

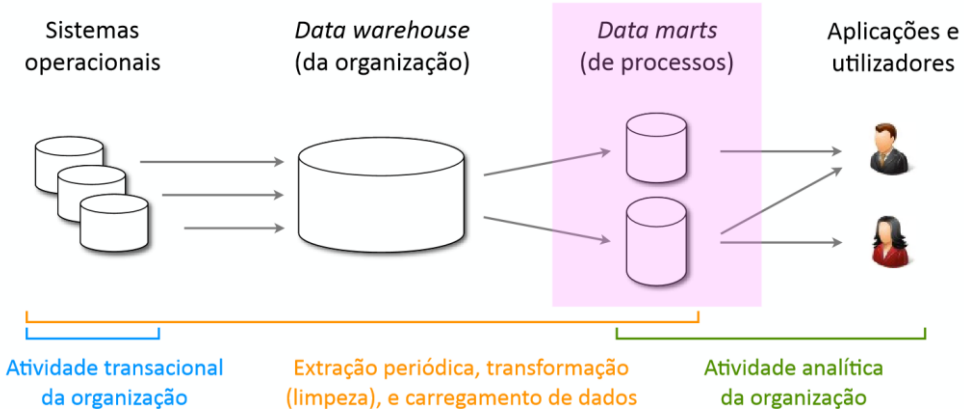


Figura 2.3 - Arquitetura de um sistema de BI visualizando *data warehouse* e *data marts* [7]

Um dos maiores desafios em sistemas de *data warehouse* é como planear a sua construção. Duas abordagens são normalmente aceites na implementação destes sistemas: uma do tipo *top-down* (Inmon) [9] e outra do tipo *bottom-up* (Kimball) [10]*:*

- *Top-down*: nesta abordagem há duas etapas, a primeira consiste na definição do modelo de dados global do *data warehouse* e a segunda etapa baseia-se na implementação de *data marts* de acordo com as necessidades e caraterísticas das várias unidades de negócio da organização. A vantagem neste tipo de implementação é ter um modelo integrador de todos os assuntos. As desvantagens são o longo tempo de implementação exigido no início, a alta taxa de risco de insucesso e a criação de expectativas em relação ao ambiente a ser construído, já que a implementação e a obtenção de resultados são demoradas.
- *Bottom-up*: neste caso**,** as partes do sistema vão sendo construídas incrementalmente. O objetivo passa por construir modelos de dados individuais, de cada *data mart*, tendo em consideração as necessidades de cada unidade de negócio. Assim os projetos são menores, focando áreas ou assuntos específicos. A estrutura do *data warehouse* é implementada de forma incremental, conforme vão sendo construídos os *data marts*. A grande vantagem é a rápida implementação, a agilidade na apresentação dos resultados, diminuindo a taxa de risco de insucesso. Esta abordagem vem-se tornando cada vez mais popular pelo facto de a abordagem *top-down* ser cara e difícil de ser definida, necessitando de mais tempo para implementação, investimento, e não apresentando um retorno rápido. A principal desvantagem da abordagem *bottom-up* é a possível dificuldade acrescida em integrar os *data marts* ao longo do tempo.

Neste trabalho foi utilizada a abordagem *bottom-up* de Kimball.

2.4 Modelação dimensional

O modelo de dados dimensional visa facilitar o acesso a grandes quantidades de dados, permitindo a análise histórica dos dados, desempenho nas consultas e facilidade no desenvolvimento das mesmas. Este modelo deve seguir uma série de requisitos: simplicidade, expressividade, precisão e eficiência [7].

- Simplicidade: os decisores devem entender o modelo e a navegação nos dados deve ser intuitiva e fácil;
- Expressividade: deve ser registada informação necessária a múltiplos cenários de decisão;
- Precisão: deve ser incluída apenas informação relevante; o máximo detalhe não significa ter dados que ninguém entende;
- Eficiência: o modelo desenhado deve permitir um desempenho eficiente; é conveniente evitar o uso excessivo de junções de tabelas para fazer relatórios.

O modelo de dados multidimensional apresenta como componentes essenciais as tabelas de dimensões e as tabelas de factos, sendo nestas últimas guardadas as medidas [7].

- As medidas são valores numéricos usados para avaliar um processo de negócio.
- As dimensões são entidades independentes que participam na análise das medidas, isto é, fornecem contexto às medidas, respondendo às perguntas quando, onde, quem e o quê? As tabelas de dimensões, têm relativamente poucas linhas, mas costumam ter muitas colunas descritivas. As tabelas de dimensões ficam ligadas entre si através das tabelas de factos.
- Os factos são eventos expressos através de dimensões e medidas. Cada tabela de factos guarda dados de eventos de um processo de negócio, possuindo chaves estrangeiras[1] (*foreign key)* que

[1] É uma referência numa tabela a uma chave primária (identificador único) de outra tabela.

referenciam as tabelas de dimensões e medidas numéricas sobre o negócio. São grandes, tipicamente com 90% da dimensão total do *data warehouse*, têm relativamente poucas colunas, e a sua atualização é sobretudo com inserções de dados e bastante mais frequente que a atualização das tabelas de dimensões.

- O grão da tabela de factos é o significado de uma linha da tabela de factos e determina o nível máximo de detalhe. O detalhe deve ser o maior possível para os recursos disponíveis. As tabelas com grão mais fino são maiores, mas também mais expressivas. O grão identifica também as dimensões e o detalhe a guardar nas mesmas. Quanto mais fino é o grão, maior tende a ser o número de dimensões. Por exemplo, considerando os alunos graduados por curso, um grão mais fino é saber, para cada curso, a nota de graduação de cada aluno; e um grão grosso é saber, por curso, quantos graduados e a média das notas. Neste último caso, deixa de existir a dimensão Aluno.

Nas tabelas de dimensões não é aconselhável usar códigos e outros identificadores do contexto do negócio como chave primária, quer por causa do desempenho, quer por causa das mudanças que possam existir nos sistemas operacionais. Assim, é habitual que se opte por uma chave substituta (*surrogate key*) sendo aconselhável que seja um número inteiro de poucos *bytes*, positivo e sequencial. Apesar de esta chave não ter nenhum significado para o processo de negócio, permite identificar cada linha da tabela de dimensão. O seu único propósito é permitir a junção da tabela de dimensão com a tabela de factos. As principais vantagens são proteger o *data warehouse* das alterações nos sistemas operacionais e simplificar a integração de dados de várias fontes. A principal desvantagem é tornar o carregamento mais complexo [8].

Existem ainda alguns conceitos relacionados com as dimensões que foram utilizados neste trabalho e são aqui mencionados:

- Dimensões degeneradas: dimensões com apenas um atributo e sem tabela associada, com significado forte no contexto e que podem servir para agrupar factos. A título de exemplo, a dimensão Opção de Candidatura, que contém exclusivamente um dígito entre 1 e 6, correspondente a uma das seis opções de candidatura de um aluno, é considerada uma dimensão deste tipo;
- Dimensões conformadas: dimensões comuns a vários processos de negócio e que partilham atributos. Por exemplo uma dimensão Curso pode ser comum aos processos de inscrição e graduação de alunos numa faculdade;
- Minidimensões: estruturas de dados que incluem diferentes combinações entre atributos, não necessariamente relacionados, que podem ser usados como critérios de pesquisa;
- Hierarquias em dimensões: permitem a análise de medidas em vários níveis de detalhe, com base em relações hierárquicas entre colunas. As operações mais habituais são as de *drill-down*, usada para procurar explicações mais detalhadas, por exemplo passar do total de diplomados por grau, para diplomados por curso, e *roll-up*, usada para aumentar o nível de agregação dos resultados (inverso do *drill-down*);
- Dimensões de mudança lenta (*slowly changing dimension*): existem dois tipos principais de alterações nas tabelas de dimensões, considerados para este trabalho:
  - Tipo 1: a nova informação é corrigida por cima da antiga, ou seja, não fica guardado o histórico, apenas a versão atual. É o típico exemplo de um erro num valor, que deve ser corrigido para melhorar a qualidade dos dados;
  - Tipo 2: reflete toda a informação em termos de histórico. Por cada mudança, é introduzida uma nova linha na tabela de dimensão, com uma data de início, uma nova chave substituta e uma data de fim de validade.

É possível que na mesma tabela de dimensão existam mudanças do Tipo 1 e do Tipo 2 e é necessário dar o seguimento adequado a cada uma, durante o processo de ETL.

Os quatro passos utilizados no procedimento de modelação dimensional são os seguintes [10]:

1) Identificar **processos de negócio** a modelar: a matriz de exequibilidade/valor (Figura 2.4) permite definir prioridades, uma vez que salienta processos de negócio valiosos e suas importâncias relativas. Esta importância deve ser avaliada em reunião com decisores e técnicos. O valor pode ser definido como o impacto de cada processo no negócio da organização, e a exequibilidade como os desafios técnicos e organizacionais para obter os dados. O processo de negócio A da Figura 2.4 tem um elevado impacto ou importância económica e também uma elevada exequibilidade;
2) Declarar o **grão** da tabela de factos;
3) Modelar as **dimensões** em tabelas;
4) Identificar as **medidas** numéricas na tabela de factos.

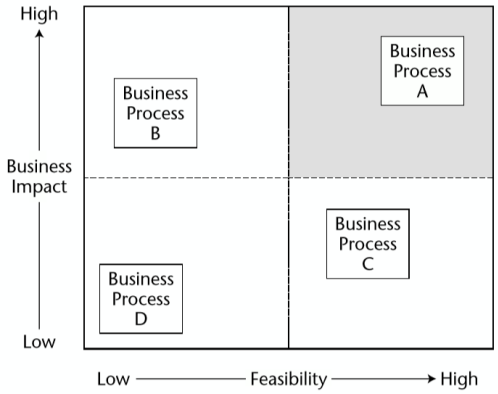


Figura 2.4 - Matriz de exequibilidade/valor [10]

O modelo de dados dimensional mais utilizado para modelar um *data warehouse* ou um *data mart* é o esquema em estrela (*star schema*), que apresenta as tabelas de dimensões em redor da tabela de factos, conforme a Figura 2.5:

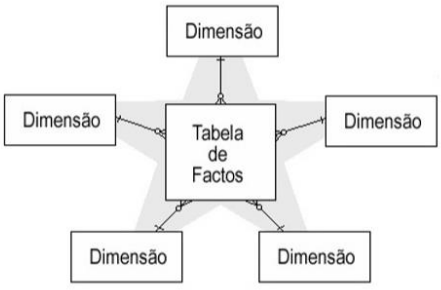


Figura 2.5 - Representação do esquema em estrela do modelo dimensional

O esquema em estrela tem vantagens para análises exploratórias: é mais fácil de entender, pois não tem muitas tabelas; ao ter poucas tabelas, o desempenho tende a ser melhor, pois há menos ligações entre tabelas. Por outro lado, é mais fácil cruzar e combinar dados, usando *drag and drop*, sendo os relatórios criados rapidamente e de forma simples, bastando arrastar atributos das dimensões e medidas numéricas dos factos. Por este motivo as tabelas de dimensões não devem ser normalizadas, uma vez que aumentaria o número de tabelas, tornaria o modelo mais complexo, as interrogações mais difíceis de escrever e mais demoradas, devido a mais junções de tabelas.

2.5 Visualização da informação

Esta secção, sobre técnicas de visualização apresenta alguns dos elementos mais utilizados para representar os dados e recomendações sobre a elaboração de relatórios/*dashboards*.

Para além de conseguir dar a resposta correta às perguntas do negócio, também é importante que esta resposta seja visualmente apelativa. Para isso existe um conjunto de opções que contribuem para obter melhores resultados: destacar determinados elementos, formatar a cor, o tamanho ou outros atributos, apresentar apenas os N elementos maiores/menores (*Top N*), ordenar de acordo com determinadas regras, aplicar filtros, aumentar ou diminuir o nível de detalhe da informação, entre outros. Atributos como o tamanho, a cor ou a posição na página são eficazes para assinalar a informação mais importante e dirigir a atenção do utilizador ao foco pretendido, criando uma hierarquia visual.

A expressão "uma imagem vale mais do que mil palavras" aplica-se também na representação de dados. A representação gráfica é uma das melhores técnicas para tornar os dados mais compreensíveis e detetar determinados padrões e relações, que de outra forma poderiam ficar ocultos. É mais fácil perceber estatísticas apresentadas em forma de gráficos de barras, diagramas de dispersão ou mapas temáticos do que numa extensa lista de números.

As melhores práticas de visualização de dados têm em conta, por um lado, a exploração dos mesmos (o que nos dizem os dados) e por outro, a sua explicação (relatar os dados aos utilizadores) [11]. Assim, existem dois fatores cruciais: o objetivo da visualização (o que se pretende transmitir) e o público-alvo.

No que diz respeito ao que se pretende transmitir, a apresentação da informação deve ser interessante e simples. A simplicidade aplica-se às tabelas, imagens e gráficos. Não se deve acrescentar elementos só porque existe essa possibilidade. Cada vez mais dispomos de um maior número de funcionalidades em cada ferramenta, por isso existe uma maior dificuldade em selecionar o que é mais apropriado [12].

No que diz respeito ao público-alvo, a informação deve ser fácil de perceber pois nem todas as pessoas estão preparadas para compreender estatísticas. Por outro lado, e uma vez que os indicadores académicos deste trabalho vão ser disponibilizados na *internet*, é preciso que o utilizador consiga encontrar de forma fácil e rápida a informação que procura. O objetivo de difundir informação através da *internet* é informar melhor o público mediante um acesso mais direto.

Quando as pessoas vão à *internet*, é muito importante captar a sua atenção. Os fragmentos devem ser pequenos para serem assimilados rapidamente. Adicionalmente, na *internet* cada secção deve ter sentido por si só. Daí a importância de a terminologia estar bem definida e os nomes das variáveis serem inteligíveis. Quando a informação é disponibilizada pela *internet*, outro aspeto fundamental é garantir a confidencialidade dos dados de pessoas individuais. Esta temática tornou-se mais relevante, desde maio de 2018, com a aplicação do Regulamento Geral sobre a Proteção de Dados [13].

Na subsecção seguinte, descrevem-se os principais elementos de visualização da informação: tabelas, gráficos (de barras ou colunas, de linhas, de dispersão) e mapas, bem como algumas práticas na sua utilização.

2.5.1 Tabelas, gráficos e mapas

As **tabelas** devem apresentar os números de forma concisa e bem organizada para ajudar na análise. De uma forma geral, as tabelas devem ser pequenas e evitar texto desnecessário. Uma casa decimal será o adequado para a maioria dos dados; só em alguns casos específicos serão precisas duas ou mais casas decimais, como para ilustrar diferenças em séries de dados. A apresentação das tabelas, classificando os dados por ordem ou por outras hierarquias, permite uma compreensão mais fácil dos dados. A ordenação também possibilita mostrar os valores mais altos e os mais baixos, assim como dados atípicos. Alinhar

à direita os valores para reforçar a sua magnitude e evitar células em branco, são também recomendações importantes.

Enquanto as tabelas interagem com o nosso sistema verbal, o que significa que as lemos, os gráficos interagem com o nosso sistema visual, que é mais rápido a processar a informação [11].

Um **gráfico** permite representar os dados de uma forma compreensível. Os gráficos podem ser muito eficazes para expressar resultados chave, devendo ter uma mensagem clara, com um título explicativo e não devendo requerer demasiado esforço para entender. Os bons gráficos estatísticos devem: ter uma imagem grande para representar muitos dados, evitar informação desnecessária ou ruído, apresentar desenhos visualmente lógicos e transmitir uma conclusão ou uma ideia [14].

Devem utilizar-se os tipos de gráficos consoante os dados: um gráfico de linhas para séries temporais, um gráfico de barras para dados ou valores absolutos ou para comparar elementos, um gráfico de dispersão para analisar relações. É necessário começar por perceber o objetivo do que se pretende transmitir para poder escolher o tipo de gráfico mais adequado. A Figura 2.6 mostra oito categorias definidas por Marco Russo [15] para agrupar representações gráficas, tendo em conta a sua finalidade ou propósito:

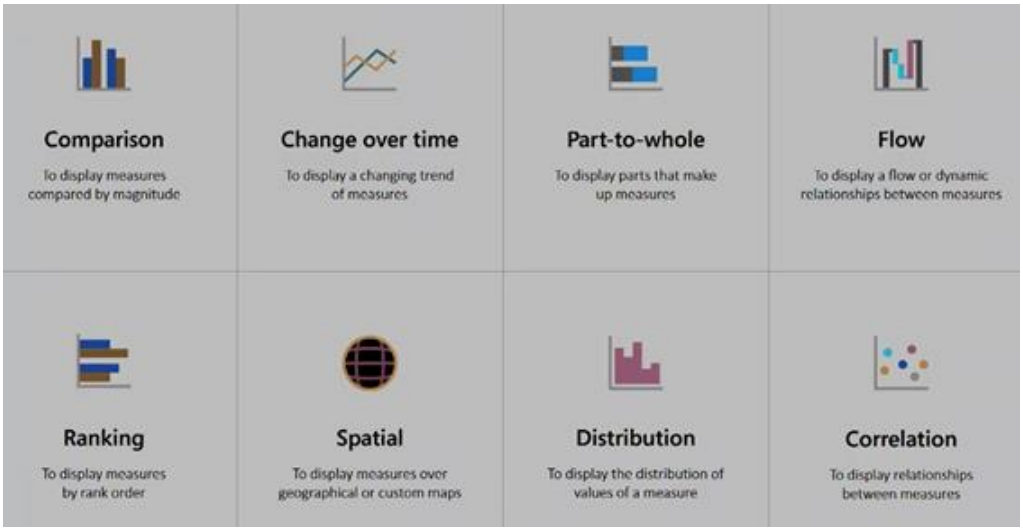


Figura 2.6 - Categorias para representações gráficas [15]

Em cada categoria podem existir diferentes tipos de representações gráficas e, inclusive, algumas representações podem pertencer a mais de uma categoria. Apesar de existirem muitos tipos de gráficos, é possível dar resposta a maioria das necessidades habituais de análise com um reduzido número deles.

Nos **mapas**, assim como nos gráficos, é fundamental saber qual o tipo de mapa mais apropriado para a informação que se pretende apresentar. Os mapas utilizam-se para ilustrar diferenças ou semelhanças entre áreas geográficas. As caraterísticas regionais ou locais, que podem ficar ocultas em tabelas ou gráficos, clarificam-se mediante a utilização de mapas. No contexto da representação dos dados, os mapas são uma área em rápida expansão, dado que os métodos de análise e representação geográfica são cada vez mais acessíveis e fáceis de utilizar [14]. O mapa mais utilizado com dados em forma de taxa é o mapa cromático, onde as diferentes tonalidades são utilizadas para mostrar contrastes entre regiões (normalmente uma cor mais escura significa um valor numérico maior). Os dados que contabilizam um parâmetro, por exemplo, total de alunos colocados por distrito, ilustram-se melhor em mapas com símbolos proporcionais. Nestes mapas, o tamanho do símbolo, como por exemplo um círculo, vai aumentando em relação ao valor numérico do dado que representa.

Estas diferentes formas de representação da informação - tabelas, gráficos e mapas - utilizam-se quer em relatórios, quer em *dashboards.*

2.5.2 Relatórios e *dashboards*

Os relatórios devem fornecer informação atual ao utilizador, informação detalhada (em vez de sumários dos dados) e flexibilidade, para permitir que os utilizadores possam criar eles próprios as suas visualizações.

Já os *dashboards* são apresentações visuais da informação mais importante, consolidada e ajustada a um único ecrã para permitir o acompanhamento rápido do negócio da organização. Os *dashboards* integram a informação de múltiplas áreas de negócio e apresentam gráficos que mostram o desempenho atual comparado com as métricas pretendidas.

Segundo Marco Russo [15] existem 15 recomendações ou boas práticas que devem ser tidas em conta na construção de relatórios ou *dashboards*, e que são apresentadas na Figura 2.7:

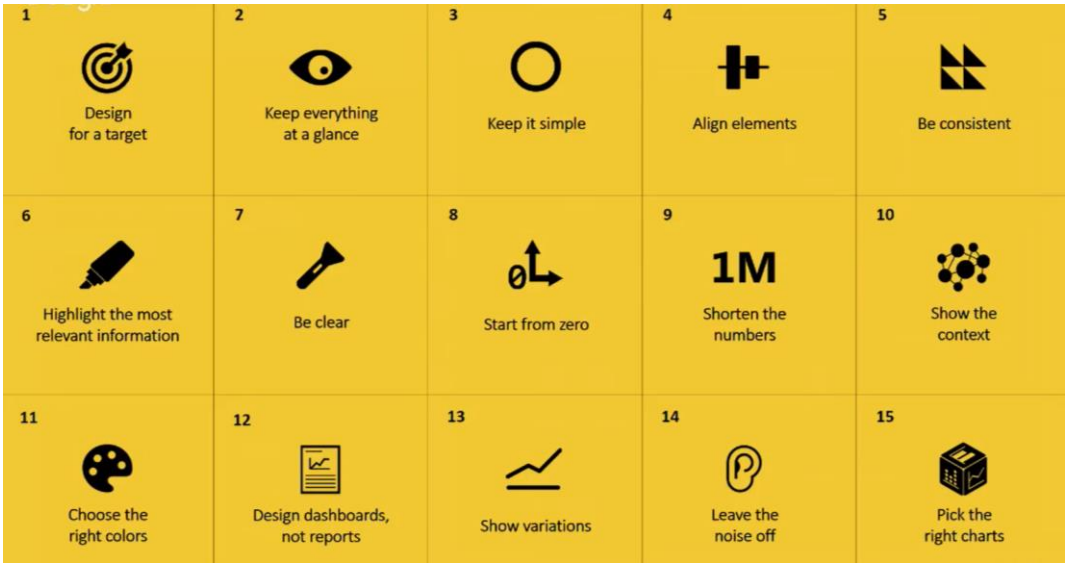


Figura 2.7 - Recomendações para a construção de relatórios e *dashboards* [15]

Cada uma das 15 recomendações é descrita de forma sucinta a seguir:

1) Definir o objetivo da visualização, isto é, saber qual a informação que se pretende mostrar no relatório ou *dashboard*;
2) Mostrar a informação num único ecrã, sem necessidade de *scroll.* Cada visualização deve ser grande o suficiente para evitar o *scroll* e pequena o suficiente para existir espaço para adicionar outra visualização;
3) Apresentar elementos simples, evitando decorações em excesso, fundos escuros, limites, isto é, elementos que não adicionem valor ao relatório e possam ser distrações;
4) Alinhar as visualizações, não só por motivos estéticos, mas para a mensagem ser mais facilmente captada;
5) Ser consistente com o uso de cores, formatação dos títulos e tipos de gráficos, e não usar diferentes tipos de visualizações apenas porque existem; usar o mesmo tipo de gráfico diversas vezes não é sinónimo de monotonia, mas pode significar profissionalismo;
6) Destacar os dados mais relevantes, pois nem todos os dados têm a mesma importância; os dados mais importantes devem estar em evidência no *dashboard*, ocupar uma área maior e estar posicionados nas zonas de maior relevância para a visão humana;

7) Ter nomes de atributos inteligíveis e inserir legendas nas visualizações, sempre que necessário; todos os símbolos ou acrónimos devem estar explícitos;
8) Começar a escala dos gráficos em zero de forma a existir um termo de comparação;
9) Formatar os números de modo a que os dados sejam mais compreensíveis; os números muito compridos são difíceis de entender;
10) Contextualizar os dados, isto é, compará-los com anos anteriores ou outros valores de referência;
11) Escolher as cores certas, pois nem todas as pessoas veem as cores da mesma maneira e certas cores ou símbolos podem ter uma conotação negativa para algumas pessoas;
12) Apresentar informação mais sumária e resumida através de *dashboards* em vez de longas listas de dados em relatórios, que não são tão fáceis de compreender;
13) Mostrar as diferenças quando elas existem, sublinhá-las e medi-las;
14) Não se deve sugerir relações que não existem; dois gráficos contíguos sem nenhuma relação devem ter uma separação suficiente para se perceber que não estão relacionados;
15) Escolher a visualização mais apropriada, por exemplo, um gráfico de barras não é a opção correta para mostrar dados evolutivos.

Uma característica habitual nos relatórios e *dashboards* é haver interatividade entre os diferentes elementos, isto é, os elementos, à medida que se vão criando, ficam vinculados e, ao interagir com algum deles, o resto dos elementos fica automaticamente alterado. Por exemplo, se selecionamos uma barra de um gráfico que apresenta um determinado ano letivo, todos os elementos desse ecrã ficam filtrados pelo mesmo ano letivo.

2.6 Ferramentas de BI

Uma vez apresentados os conceitos teóricos, incluindo os fundamentos da modelação e visualização de dados, é necessário fazer a escolha da ferramenta de BI a utilizar neste projeto. Para isso, é necessário responder à seguinte pergunta: que ferramentas e tecnologias se adaptam melhor ao perfil de quem vai disponibilizar a informação, de quem a vai consumir e das necessidades da Faculdade de Ciências? [5]

Hoje em dia existe uma grande variedade de ferramentas e sistemas de BI, que estão em permanente processo de evolução e melhoria devido ao elevado crescimento que se está a produzir neste mercado e às previsões de que continuará a ser assim nos próximos anos. De acordo com a consultora Gartner [16], foram consideradas como líderes de mercado, em 2018, as empresas Microsoft, Tableau e Qlik (ver Figura 2.8).

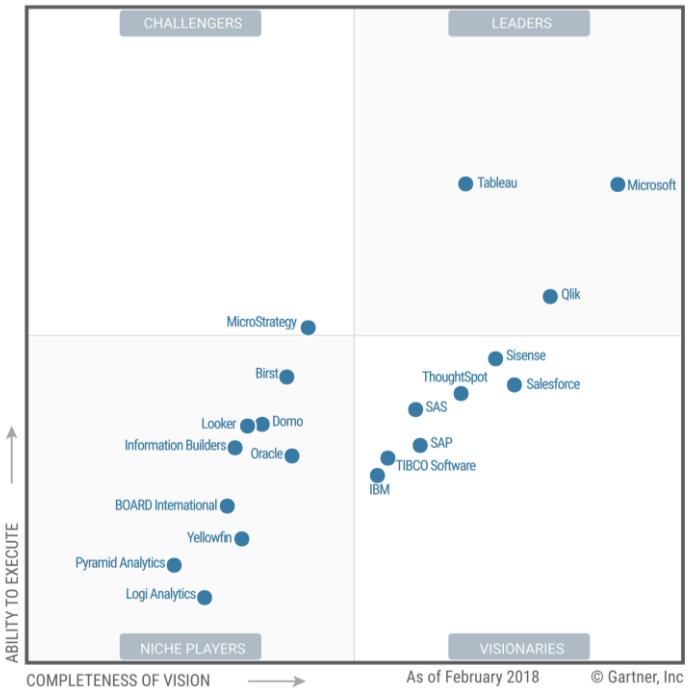


Figura 2.8 - Gartner Magic Quadrant for BI & Analytics Platforms [16]

A Figura 2.8 posiciona as empresas em quatro categorias: líderes, aspirantes, visionários e nichos de mercado. Desta forma, e uma vez que esta análise é realizada anualmente, é possível ver a evolução de uma empresa e o seu posicionamento em relação aos concorrentes. Desde o ano de 2008 que a Microsoft se posiciona nos líderes no *Magic Quadrant BI*.

Tendo em conta as caraterísticas e objetivos deste projeto, nomeadamente existir autonomia por parte do utilizador de negócio em selecionar os relatórios mais adequados e ser uma ferramenta sem custos para a faculdade, a plataforma escolhida foi a da Microsoft: Power BI. Contudo, a Microsoft possui outra plataforma de BI popular e completa, inserida no SQL Server, mas tendo em conta o requisito do *self-service* BI, o Power BI torna-se mais apropriado para este trabalho.

Um outro estudo realizado pela Stratebi [17], em abril de 2017, que analisa e compara sete das ferramentas de BI com maior aceitação atualmente no mercado (PowerBI, Tableau, Qlikview, Pentaho, SAS, Information Builders, Amazon Quicksight) atribui a pontuação máxima ao Power BI em termos de criação de *dashboards*, multiplataforma e preço, e também confere uma pontuação elevada nas componentes de transformação/modelação de dados, funcionalidade analítica e visualizações.

2.7 Visão geral do Power BI

O Power BI é um conjunto de serviços de *software*, aplicações e ligações que transformam dados em análises interativas, sem depender de uma equipa de Tecnologias de Informação (TI). Trata-se de uma ferramenta de integração, análise e visualização de dados, com o objetivo de extrair os dados de diversas fontes e criar modelos que ajudem a analisar o negócio [18].

A primeira versão do Power BI, baseada nos suplementos do Microsoft Excel (Power Query, Power Pivot e Power View) surgiu em setembro de 2013. Em julho de 2015 a Microsoft lançou uma ferramenta em que os três suplementos já estavam integrados numa única aplicação: o Power BI.

A tecnologia utilizada no Power BI, que realiza todos os cálculos em memória, é a denominada xVelocity, que permite a compressão de dados.

2.7.1 Componentes principais e blocos de construção básicos

Os componentes principais do Power BI são mostrados na Figura 2.9. O Power BI Desktop é utilizado para integrar dados, criar modelos analíticos e gerar relatórios. O Power BI Service tem como objetivo publicar, criar *dashboards* e partilhá-los de forma segura em cada organização/instituição. O Power BI Mobile é utilizado para exibir e interagir através de dispositivos móveis.

Neste projeto foi utilizado essencialmente o Power BI Desktop, cujos ficheiros têm uma extensão .pbix. O Power BI Service foi utilizado exclusivamente para a parte da publicação dos relatórios na *web*.

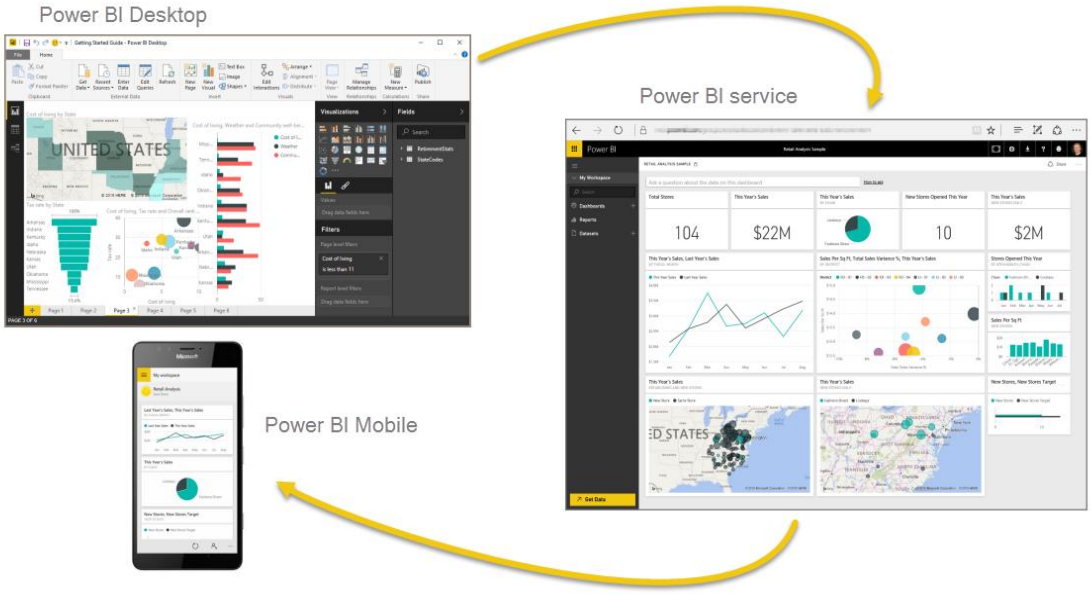


Figura 2.9 - Componentes do Power BI [18]

O fluxo comum de trabalho no Power BI é criar um relatório no Power BI Desktop, publicá-lo no Power BI Service e partilhá-lo com outras pessoas para que elas possam interagir através do Power BI Service ou numa aplicação móvel (Power BI Mobile).

O Power BI Desktop é uma aplicação para Windows que corre localmente num computador pessoal, enquanto o Power BI Service é um serviço na nuvem que se usa através do navegador da *web* [19].

O Power BI está estruturado em blocos de construção básicos, que são os seguintes:

- Visualizações: representação visual dos dados, como um gráfico ou um mapa. As visualizações podem ser simples – como um único número que representa algo significativo – ou ser visualmente complexas – como um mapa de cores de gradiente. O objetivo de uma visualização é apresentar dados de modo a fornecer contexto e informações, que, provavelmente, seriam difíceis de serem assimilados numa tabela ou num texto. O Power BI tem uma grande variedade de visualizações disponíveis, conforme se pode verificar na Figura 2.10, desde gráficos de barras simples, gráficos de dispersão e mapas até outros menos usuais, como gráficos de funil, gráfico de cascata e medidores. Após a elaboração da visualização pretendida, é possível redimensioná-la, movê-la ou alterar o tipo de gráfico, consoante a necessidade. Também é possível fazer formatações sobre o texto do título e cores de dados, através da seleção do ícone de pincel no painel Visualizações (ver Figura 2.10):

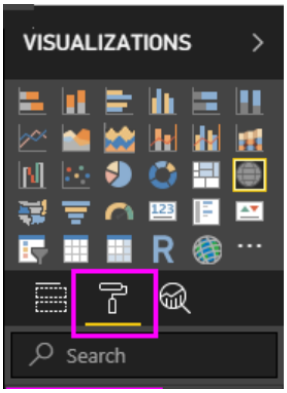


Figura 2.10 - Tipos de visualização no Power BI Desktop [18]

- Conjuntos de dados: coleção de dados que o Power BI usa para criar visualizações. Pode ser simples (baseado numa única tabela de um ficheiro Excel) ou uma combinação de várias fontes diferentes. Por exemplo, pode criar-se um conjunto de dados baseado em três bases de dados, numa tabela de um *site* e numa tabela do Excel.
- Relatórios: coleção de visualizações, relacionadas entre si, que aparecem juntas numa ou mais páginas. Os relatórios permitem organizar as visualizações de diversas maneiras. Pode existir, por exemplo, um relatório sobre alunos colocados, um relatório sobre diplomados e um relatório sobre os cursos.
- Painéis (*dashboard*): grupo selecionado de visualizações que fornecem uma análise rápida dos dados e que deve ser ajustado a uma única página, geralmente chamada de ecrã. Quando se partilha uma única página de um relatório ou uma coleção de visualizações, trata-se de um *dashboard*.

No Power BI os relatórios podem ser publicados numa página *web*, enquanto que os *dashboards* podem ser partilhados internamente na organização. Nos *dashboards* as visualizações não são interativas como nos relatórios. Tendo em conta os objetivos deste trabalho, disponibilização de indicadores académicos na página *web* da faculdade e interação das visualizações, foram utilizados os relatórios e não os *dashboards*.

2.7.2 Power BI Desktop - Editor de Consultas

A ferramenta do Power BI para formatação e transformação de dados, de modo a que fiquem prontos para a modelação e visualização, é o Editor de Consultas. Trata-se de uma ferramenta ETL destinada ao utilizador do negócio. Este editor, que abre numa janela independente da interface de utilizador do Power BI Desktop, está dividido nas seguintes secções que podem ser visualizadas na Figura 2.11:

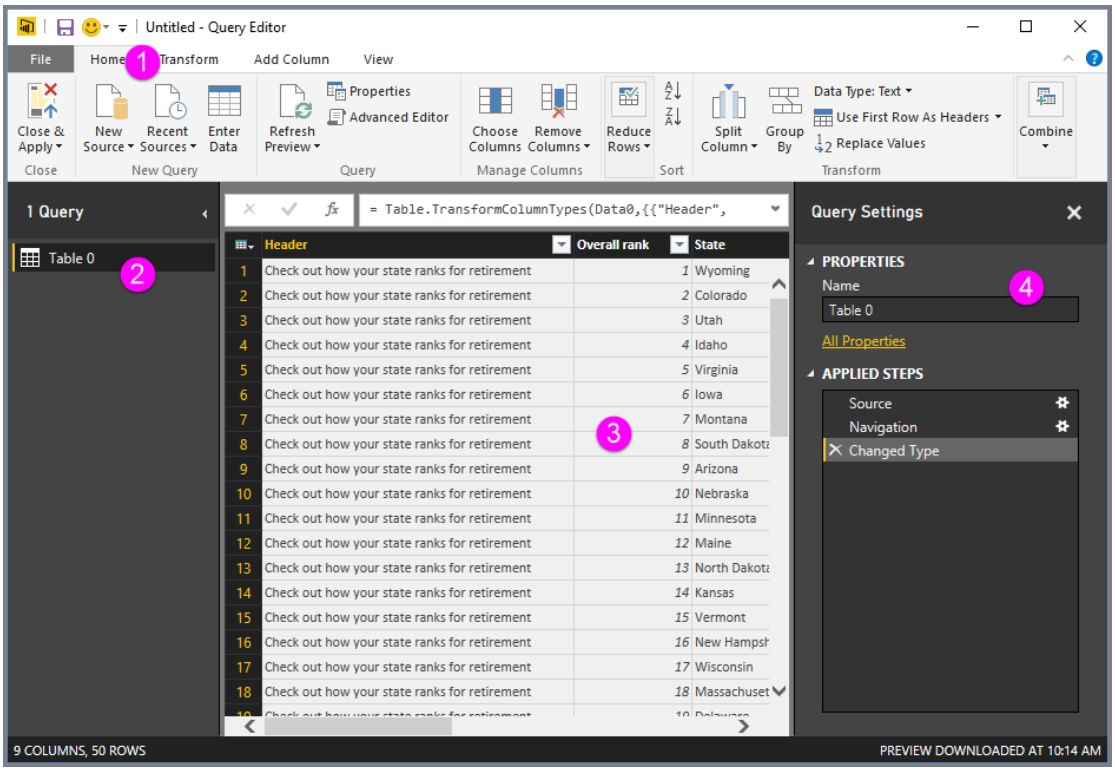


Figura 2.11 - Editor de consultas do Power BI Desktop [18]

1) Os menus e respetivos comandos para interagir com os dados na consulta;
2) No painel esquerdo, as consultas (uma para cada tabela) são listadas e ficam disponíveis para seleção e formatação;
3) No painel central constam os dados da consulta selecionada, disponíveis para formatação;
4) O painel de Configurações de Consulta (*Query Settings*) mostra as propriedades da consulta e as etapas aplicadas; à medida que se aplicam as transformações, cada passo é mostrado na lista Passos Aplicados (*Applied Steps*), no lado direito do Editor de Consultas.

Cada um dos passos que vão sendo executados em relação à obtenção de dados e à sua posterior transformação são registados e geram internamente expressões em linguagem M, código este que pode ser acedido no Editor Avançado.

2.7.3 Power BI Desktop – Modos de Relações, Dados e Relatório

Após a obtenção, transformação e carregamento dos dados, é possível, através do Modo de **Relações,** a visualização gráfica das relações entre tabelas, e a sua eventual correção, de acordo com as relações estabelecidas no modelo de dados. De seguida, realiza-se no Modo de **Dados,** a criação de hierarquias de dimensões e o enriquecimento do modelo de dados, através da criação de novas medidas. Por último, é executada, no Modo de **Relatório,** a componente de análise (criar relatórios, adicionar filtros, escolher visualizações).

O menu principal do Power BI Desktop, no Modo de Relatório, tem cinco áreas fundamentais, conforme a Figura 2.12:

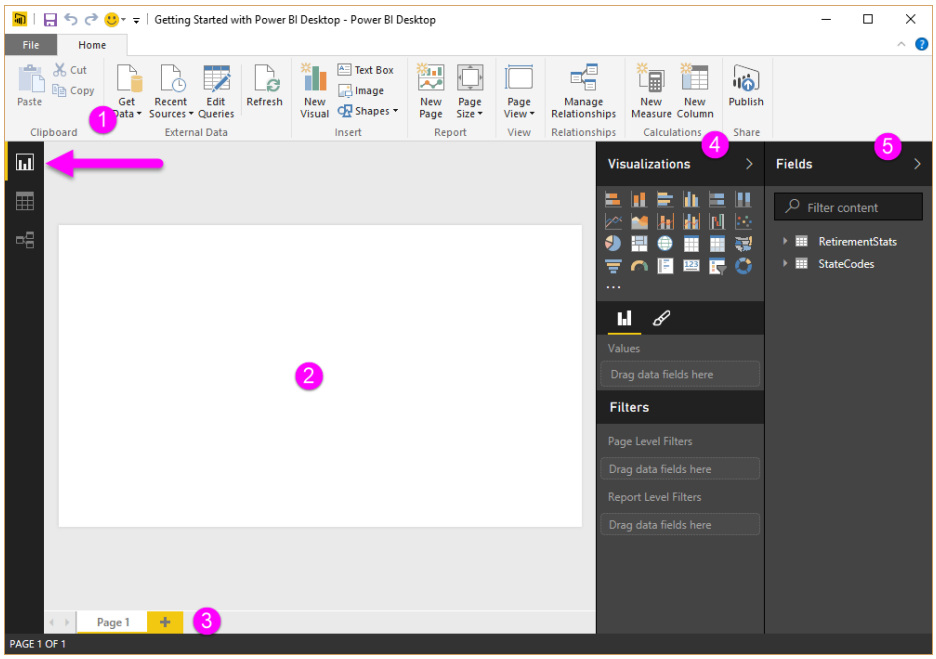


Figura 2.12 - Modo de Relatório do Power BI Desktop [18]

1) O painel de opções, que permite alternar entre os modos de Relatório (selecionado), Dados e Relações;
2) O ecrã do Relatório, em que as visualizações são criadas e organizadas;
3) A área das Páginas na parte inferior, que permite selecionar ou adicionar uma página ao relatório;
4) O painel Visualizações (*Visualizations*), em que é possível alterar visualizações, personalizar cores ou eixos, aplicar filtros e arrastar campos;
5) O painel Campos (*Fields*), em que constam as várias tabelas e atributos dos dados existentes, bem como as medidas que tenham sido geradas. Ao fazer-se clique sobre o triângulo à esquerda do nome de uma tabela, torna-se possível ver atributos, medidas e outros componentes da mesma.

Existe ainda um painel de filtros que, como o nome indica, permite filtrar os dados a apresentar no relatório. Estes filtros podem ser considerados ao nível da visualização ativa, ao nível da página ou ao nível de todo o relatório.

Conforme referido anteriormente, para enriquecer o modelo é necessário executar cálculos mais complexos de modo a criar medidas próprias. Para isso utiliza-se a linguagem DAX (*Data Analysis eXpressions*). O DAX é uma linguagem de fórmulas que incrementa a capacidade de cálculo analítico e a velocidade de resposta. O DAX usa muitos operadores, funções e sintaxe também utilizados nas fórmulas do Excel. No entanto, as funções DAX foram projetadas para trabalhar com dados relacionais e realizar cálculos mais dinâmicos. Há mais de 200 funções DAX que fazem, desde agregações simples, como soma e média, até funções de estatística e de filtragem mais complexas. O DAX pode ser utilizado para criar dois cálculos principais:

- Colunas calculadas: colunas adicionais inseridas no modelo de dados. São cálculos de cada linha na tabela, utilizados para segmentar ou filtrar. As colunas calculadas estão também armazenadas, isto é, contam para o peso do modelo. Um exemplo de uma coluna calculada é a idade de um aluno à data da conclusão do curso, com base na sua data de nascimento.
- Medidas calculadas: são cálculos definidos, como agregações, somas, rácios, contagens, médias, percentagens ou taxas. Só podem ser utilizadas na área dos valores e não das colunas

ou linhas. A grande vantagem é que não ocupam espaço no modelo. As medidas calculadas podem ser usadas como um argumento em outras fórmulas, que torna as fórmulas e o modelo mais eficientes. Um exemplo de uma medida calculada é a Média da Classificação Final.

A partir da versão de fevereiro de 2018 do Power BI Desktop, muitos cálculos, como médias, mínimos, máximos, adições, subtrações, entre outros, estão disponíveis como medidas rápidas (cerca de 25). São medidas pré-definidas e que podem ser selecionadas através de uma caixa de diálogo. Estes cálculos rápidos também permitem aprender a sintaxe DAX, uma vez que as respetivas fórmulas DAX estão disponíveis para análise.

2.8 Sumário

Neste capítulo apresentaram-se conceitos teóricos relevantes num sistema de BI, incluindo fundamentos da modelação de dados e de técnicas de visualização. Estes conceitos são aplicáveis às diferentes ferramentas disponíveis no mercado, tendo sido descrito com mais detalhe o Power BI, ferramenta escolhida para este projeto. Estes conceitos teóricos foram aplicados no Capítulo 4, que descreve a concretização da solução.

3. Análise do problema

Neste capítulo são descritos os requisitos de alto nível que caraterizam o problema a resolver, são identificados os processos de negócio para o projeto, é feita uma descrição sobre as fontes de dados utilizadas no trabalho, são definidos os indicadores académicos e, por último, são descritos os procedimentos e as metodologias que eram utilizados pelos colaboradores do Gabinete de Avaliação e Auditoria Interna (GAAI) da Faculdade de Ciências (FCUL).

3.1 Definição de requisitos gerais

A definição de requisitos é um dos pontos cruciais no desenvolvimento de qualquer tipo de projetos e passa por identificar as necessidades dos diferentes utilizadores, bem como da organização no seu todo, de modo a que o sistema a construir venha a dar a resposta adequada. A identificação das necessidades tem vindo a ser feita ao longo dos anos com base nos pedidos remetidos ao GAAI. A Figura 3.1 apresenta um resumo dos principais instrumentos solicitados ao GAAI pelas diferentes partes interessadas e às quais é necessário dar uma resposta em termos de indicadores, atempada e com a informação correta:

Órgãos de Gestão	Reporte Institucional	Comunidade FCUL
• **Direção de Ciências**: Instrumentos de gestão estratégica (Plano/Relatório de Atividades); Instrumentos do Sistema Integrado de Garantia da Qualidade	• **Agência de Avaliação e Acreditação do Ensino Superior (A3ES)**: Processos de Avaliação/Acreditação de ciclos de estudos e Avaliação Institucional • **Reitoria da ULisboa**: Instrumentos de gestão universitária (Plano/Relatório de Atividades) • ***Rankings* Internacionais**: U-Multirank; THE; QS • **Acreditações:** Nacionais (Ordem dos Engenheiros Técnicos) e Internacionais (EUR-ACE)	• **Departamentos e Unidades de Serviço:** Instrumentos de gestão interna (Plano/Relatório de Atividades; • Relatório do Departamento; • Relatório do Ciclo de Estudos); • Eventos (Dia de Ciências); • Prémios de desempenho e distinções; • Brochuras, Folhetos e Media (Agenda de Ciências, Revista Forum Estudante, Guias do Estudante do Semanário Expresso)

Figura 3.1 - Principais instrumentos de gestão solicitados ao GAAI, em termos de indicadores

Por outro lado, uma vez que esta informação também é necessária para departamentos e unidades de serviço da FCUL, torna-se essencial a sua disponibilização através do Portal de Ciências. Assim, e de modo a facilitar esta visualização da informação, foram identificados os seguintes requisitos gerais:

1. **Relatórios interativos**: pretende-se que, ao selecionar um determinado atributo numa visualização, as restantes visualizações da página do relatório sejam filtradas por esse atributo;

2. **Existência de filtros**: um dos grandes objetivos em termos de requisitos é que a obtenção da informação possa ser realizada pelo próprio utilizador, através de diferentes escolhas. A título de exemplo:
    a. Escolha de um determinado ano letivo ou visão evolutiva;
    b. Escolha de um determinado curso, dos vários cursos de um determinado departamento ou de todos os cursos de um determinado grau;
3. **Existência de hierarquias**: a obtenção da informação possa ser realizada pelo próprio utilizador com maior ou menor nível de detalhe, através das opções de *drill down* e *roll up;*
4. **Publicação na *web***: pretende-se que estes relatórios sejam disponibilizados no Portal de Ciências (https://ciencias.ulisboa.pt/pt/estatísticas);
5. **Exportação dos relatórios para PDF**: pretende-se ainda que o conteúdo destes relatórios seja exportável para PDF;
6. **Exportação dos dados para ficheiros**: os dados que dão origem às visualizações também devem poder ser exportados.

Adicionalmente a estes requisitos gerais, devem ser cumpridos os requisitos transversais a qualquer *data warehouse*, mencionados na Secção 2.3. É também necessário que o sistema desenvolvido cumpra as recomendações sobre a construção de relatórios, referidas na Subsecção 2.5.2, e dê resposta às perguntas de negócio, identificadas de seguida.

3.2 Processos de negócio académicos

Para além da definição de requisitos gerais, outro dos fatores cruciais do sucesso da implementação de um sistema de *Business Intelligence* (BI) tem a ver com o conhecimento dos processos de negócio. É fundamental trabalhar de perto com os responsáveis pela tomada de decisão, e com os utilizadores de negócio implicados no processo final. Só assim se conseguem identificar as necessidades passadas, presentes e futuras, de modo a que o novo sistema as resolva da melhor maneira possível e consiga responder às perguntas analíticas identificadas.

Os pilares de uma Instituição de Ensino Superior (IES), por regra, são os seguintes:

- Ensino (Formação)
- Investigação, Desenvolvimento e Inovação
- Transferência de Conhecimento

Estes três pilares estão assentes em Recursos (Humanos, Físicos/Materiais, Financeiros).

De acordo com a Lei n.º 62/2007, que estabelece o regime jurídico das instituições de ensino superior [20], as IES aprovam e publicam um relatório anual, dando conta, designadamente: da evolução das admissões de alunos e da frequência dos ciclos de estudos ministrados; dos graus académicos e diplomas conferidos; da empregabilidade dos seus diplomados; da internacionalização da instituição e do número de estudantes estrangeiros.

Neste contexto, podem identificar-se os seguintes processos de negócio, relacionados com a temática do Ensino: Acesso, Inscrição, Conclusão, Empregabilidade e Internacionalização.

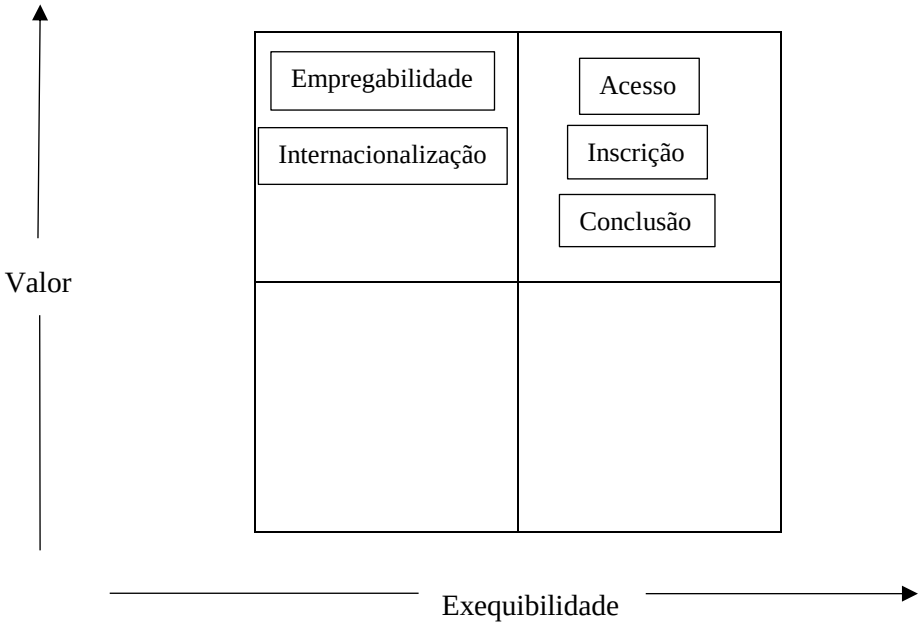


Figura 3.2 - Matriz de exequibilidade/valor deste projeto

A matriz de exequibilidade/valor, apresentada na Figura 3.2, salienta os processos de negócio prioritários. Tendo em conta que os dados relativos à empregabilidade e à internacionalização não são obtidos no GAAI, sendo da responsabilidade de outra unidade de serviço da FCUL, ficaram posicionados no quadrante de alto valor, mas exequibilidade mais baixa. Assim, os três processos considerados neste trabalho, foram os seguintes:

- **Acesso**: processo que tem como objetivo analisar a procura da Faculdade de Ciências por parte dos alunos candidatos ao Ensino Superior, através do Concurso Nacional de Acesso (CNA). Neste contexto, são caraterizados os alunos:
    - **Candidatos -** todos aqueles que, nos termos do regulamento do CNA, instruem a candidatura a pelo menos uma das fases, através do preenchimento do formulário *online* [21];
    - **Colocados –** todos aqueles que, no final do processo de candidatura e após a seriação, se encontram numa situação que lhes confere o direito à matrícula e inscrição no par instituição/curso a que se candidataram numa determinada fase do CNA [21];
    - **Matriculados –** todos aqueles que efetivaram a sua matrícula na instituição e inscrição no curso em que foram colocados.
- **Inscrição**: processo que pretende caraterizar os alunos **inscritos** em Ciências. Considera-se inscrito "todo aquele que, no dia 31 de dezembro do ano N, está inscrito num estabelecimento de ensino superior, num curso nele lecionado" [22];
- **Conclusão**: processo que tem como objetivo analisar a eficiência formativa e caraterizar os alunos **diplomados** de Ciências. Considera-se diplomado: "todo o aluno que entre 1 de janeiro e 31 de dezembro do ano N reuniu as condições legalmente previstas para a emissão do diploma de um dos níveis de formação, independentemente de ter ou não solicitado a sua emissão" [22].

Os alunos inscritos e diplomados considerados neste trabalho são provenientes do inquérito anual destinado a todas as IES, denominado Registo de Alunos Inscritos e Diplomados no Ensino Superior (RAIDES) e descrito detalhadamente na Secção 3.3.

Existem outros processos de ensino relacionados com o desempenho académico (sucesso escolar) mas que também não foram considerados neste trabalho, uma vez que a sua obtenção depende do sistema académico (FenixEdu), que se encontra em fase de estabilização.

Algumas das perguntas de negócio às quais este trabalho pretendeu dar resposta, encontram-se identificadas a seguir, agrupadas por processo de negócio:

Processo de negócio: Acesso

- Qual a evolução da taxa de ocupação dos cursos de licenciatura e mestrado integrado?
- Qual o índice de satisfação da procura[2] de um determinado curso, num dado ano letivo?

Processo de negócio: Inscrição

- Qual tem sido a tendência do número total de inscritos de Ciências, por nível de formação?
- Qual a percentagem de novos alunos de mestrado provenientes de outra instituição?

Processo de negócio: Conclusão

- Quais os cursos de licenciatura em que os alunos se graduam com uma média de classificação final mais elevada?
- Qual a média do número de anos até à conclusão dos cursos, por nível de formação?

A Figura 3.3 apresenta os processos de negócio, as respetivas fontes de dados e a tipologia do aluno em cada situação:

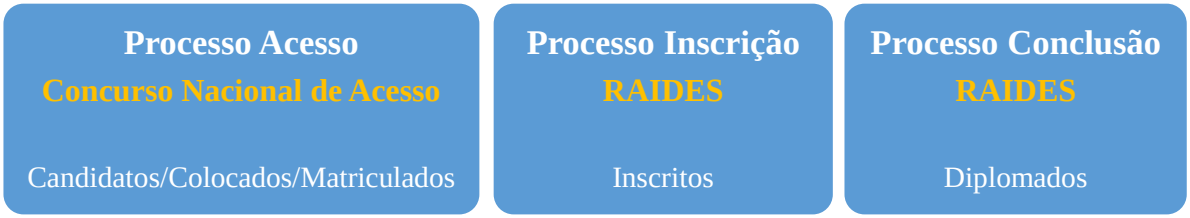


Figura 3.3 - Processos de negócio académicos considerados neste trabalho e respetivas fontes de dados

A Secção 3.3 apresenta uma descrição detalhada de cada uma das fontes de dados utilizadas neste trabalho.

3.3 Fontes de dados

Os dados utilizados neste projeto provêm essencialmente de duas fontes:

- Dados sobre os resultados do Concurso Nacional de Acesso (CNA): obtidos através da Direção-Geral do Ensino Superior (DGES). O CNA é a principal forma de acesso ao Ensino Superior Público e representa cerca de 85% dos novos estudantes de licenciaturas e mestrados integrados, no caso das instituições universitárias. Os dados sobre os restantes regimes de acesso, nomeadamente os concursos especiais, não foram considerados neste trabalho;
- Dados sobre o Registo de Alunos Inscritos e Diplomados no Ensino Superior (RAIDES): obtidos através da Direção-Geral das Estatísticas de Educação e Ciência (DGEEC). O RAIDES é um inquérito anual, de âmbito nacional, dirigido a todos os estabelecimentos do ensino

[2] O índice de satisfação da procura é o quociente entre o número de candidatos em 1ª opção e o número de vagas.

superior, que visa caraterizar o sistema de ensino superior, na vertente de alunos inscritos e diplomados [23].

Em termos de horizonte temporal, os dados utilizados neste trabalho dizem respeito ao período compreendido entre o ano de 2013 e o ano de 2017: os dados do CNA iniciam-se com o CNA13 (ingresso no ano letivo 2013/14) e os dados do RAIDES com o RAIDES13 (inscritos 2013/14 e diplomados 2012/13).

O ano de 2013 foi escolhido como ano letivo de início deste trabalho pelos seguintes motivos:

- A transição do RAIDES12 para o RAIDES13, teve mudanças significativas em diversas vertentes. Até ao RAIDES12 a exportação era realizada através de uma base de dados Access e a partir do RAIDES13 passou a ser feita em formato XML (eXtensible Markup Language). Por outro lado, um número significativo de variáveis foi adicionado e outras foram removidas. As próprias codificações dos valores de algumas variáveis sofreram alterações.
- A opção de não considerar os cursos pré-bolonha: a FCUL teve alunos inscritos e diplomados de cursos pré-bolonha até ao ano letivo 2010/11;
- As regras do RAIDES sofreram diversas alterações no que diz respeito aos diplomas de especialização dos cursos de mestrado e de doutoramento, que apenas estabilizaram a partir do ano letivo 2012/13;
- As regras do RAIDES tiveram diversas alterações no que diz respeito aos alunos de mobilidade (*incoming*): estes alunos eram incluídos como alunos inscritos e apenas obtiveram um estatuto autónomo denominado mobilidade a partir do ano letivo 2012/13.

A seguir são mostrados os detalhes sobre cada fonte de dados.

3.3.1 Concurso Nacional de Acesso

O CNA é o processo destinado à colocação dos candidatos à matrícula e inscrição no ensino superior público, em cada ano letivo, nas vagas existentes para cada par estabelecimento/curso.

A inscrição é o ato administrativo que faculta, depois de efetivada a matrícula (ato de registo), a frequência de um determinado ano escolar, disciplina ou curso.

Os resultados dos candidatos colocados através do Concurso Nacional de Acesso são disponibilizados às IES por via eletrónica, através das respetivas áreas reservadas no endereço https://pontounico.dges.gov.pt, com credenciais de acesso geradas e fornecidas pela DGES. A base de dados disponibilizada às faculdades da Universidade de Lisboa contém a informação dos alunos candidatos apenas a esta Universidade. No âmbito deste projeto foram analisados os dados relativos aos alunos candidatos e/ou colocados na FCUL na 1ª fase do CNA.

A referida base de dados, em Microsoft Access, com informação sobre candidatos e colocados contém cerca de 30 tabelas, tendo para este trabalho sido utilizadas as que constam na Tabela 3.1:

Tabela 3.1 - Tabelas dos dados provenientes do CNA utilizadas neste trabalho

Nome da tabela	Descrição	Nº de linhas (em 2017)	KB
tblAlunosCand	Dados gerais sobre a candidatura na fase a que respeita a base de dados, por candidato, incluindo contingente, etapa de colocação, par instituição/curso de colocação	21609	1507

tblAlunos	Data de nascimento, sexo, dados sobre o ensino secundário	21609	2578
tblAlunosIdent	Identidade e contactos dos candidatos	21609	1887
tblAlunosPrefer	Preferências de par instituição/curso dos candidatos	112137	9929
tblEstSup	Instituições de Ensino Superior e respetivas unidades orgânicas	299	23
tblCursup	Ciclos de estudos do ensino superior	702	31
tblEscolas	Estabelecimentos de ensino secundário	652	37
tblCodsEtapCol	Etapas de colocação (Contingente geral, Açores, Madeira, Emigrantes, Militar ou Deficiente)	15	1
tblCodsPais	Países de proveniência	249	8
tblCodsDistrito	Distritos e regiões autónomas do território nacional	21	1
tblAnoActual	Ano da candidatura a que respeita a base de dados	1	1

O número de linhas e o número de *bytes* dizem respeito ao ficheiro dos dados relativos à 1ª fase do CNA17.

A Figura 3.4 representa, em vista de diagrama, as relações entre as principais tabelas anteriores:

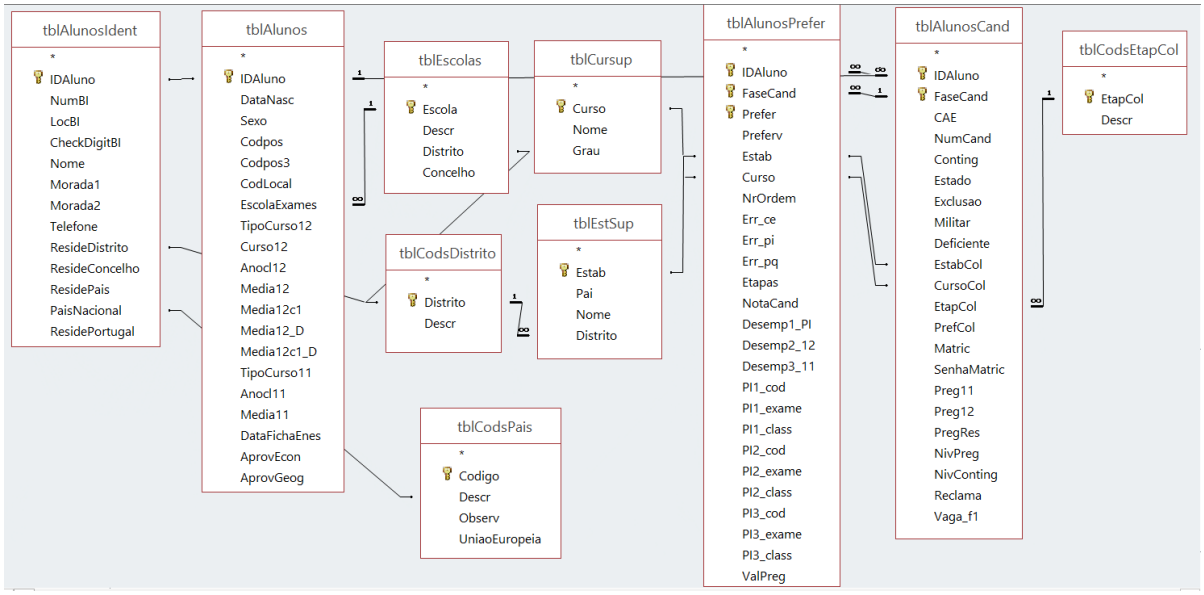


Figura 3.4 - Relações entre as tabelas dos dados provenientes do CNA utilizadas neste trabalho

Conforme se pode constatar na Figura 3.4, os atributos têm nomes pouco inteligíveis. Em relação às tabelas Alunos, Alunos Identidade, Alunos Preferência e Alunos Candidatos, apenas foi utilizado, neste trabalho, um subconjunto dos atributos disponíveis, conforme apresentado na Figura 3.5:

tbAlunos

Nome do campo	Tipo de dados	Descrição (Opcional)
IDAluno	Número	Identificador de aluno
DataNasc	Data/Hora	Datade nascimento
Sexo	Texto Breve	Sexo do aluno
EscolaExames	Texto Breve	Código de escola de realização de exames

tbAlunosIdent

Nome do campo	Tipo de dados	Descrição (Opcional)
IDAluno	Número	Identificador de aluno
NumBI	Número	Nº de identificação (BI/CC/...)
LocBI	Texto Breve	Tipo de documento de identificação
Nome	Texto Breve	Nome do aluno
ResideDistrito	Texto Breve	Distrito de residência à data da candidatura
ResideConcelho	Texto Breve	Concelho de residência
PaisNacional	Texto Breve	País de nacionalidade

tbAlunosPrefer

Nome do campo	Tipo de dados	Descrição (Opcional)
IDAluno	Número	Identificador de aluno
FaseCand	Texto Breve	Fase de candidatura
Preferv	Texto Breve	Preferência válida
Estab	Texto Breve	Codigo de instituição
Curso	Texto Breve	Codigo de curso
NotaCand	Número	Nota de candidatura

tblAlunosCand

Nome do campo	Tipo de dados	Descrição (Opcional)
IDAluno	Número	Identificador de aluno
FaseCand	Texto Breve	Fase de candidatura
Conting	Texto Breve	Contigente pelo qual se candidatou
EstabCol	Texto Breve	Instituição de colocação
CursoCol	Texto Breve	Curso de colocação
EtapCol	Número	Etapa de colocação
PrefCol	Texto Breve	Preferência válida de colocação
Matric	Texto Breve	Efectuou matrícula S/N

Figura 3.5 - Atributos, das principais tabelas do CNA, utilizados neste trabalho

Os dados analisados neste trabalho englobam cinco anos letivos. A Tabela 3.2 apresenta os valores relativos ao número de candidaturas, ao número de colocados e ao número de matriculados em Ciências, por ano letivo.

Tabela 3.2 - Dados provenientes do CNA relativos aos últimos cinco anos letivos

Ano letivo	2013/14	2014/15	2015/16	2016/17	2017/18
Nº de Candidaturas	3459	3181	4415	4813	5423
Nº de Colocados	766	665	819	892	924
Nº de Matriculados	702	600	744	814	851

Nos cinco anos em análise, ocorreram as seguintes alterações nas tabelas provenientes do CNA e com implicações neste trabalho:

- CNA16: A tabela das Entidades emissoras do documento de identificação, passou de 9 valores possíveis (Civil, Macau, Exército, Marinha, Força Aérea, P.S.P, G.N.R., Guarda-Fiscal e Nº

Interno) para apenas 2 (Civil e Nº Interno). O número interno é um número atribuído anualmente pela DGES quando o aluno não tem Cartão de Cidadão/Bilhete de Identidade;

- CNA14: Foi introduzida uma nova tabela com Códigos de Países (tblCodsPais);
- CNA14: Na tabela de Identificação dos Alunos (tblAlunosIdent), foram introduzidos os seguintes cinco novos atributos:

Nome do campo	Tipo de dados	Descrição (Opcional)
ResideDistrito	Texto Breve	Distrito de residência à data da candidatura
ResideConcelho	Texto Breve	Concelho de residência
ResidePais	Texto Breve	País de residência, se não for Portugal
PaisNacional	Texto Breve	País de nacionalidade
ResidePortugal	Texto Breve	Tem residência em Portugal

Figura 3.6 - Atributos introduzidos na tabela de Identificação dos Alunos, em 2014

No Capítulo 4 descreve-se qual a metodologia utilizada para dar resposta às alterações mencionadas.

3.3.2 Registo de Alunos Inscritos e Diplomados no Ensino Superior

O RAIDES tem como unidade estatística de observação o aluno e, conforme referido anteriormente, tem por objetivo a caraterização do sistema de ensino superior, na vertente de alunos inscritos e diplomados.

A informação recolhida através deste inquérito responde à obrigatoriedade da divulgação pública anual de estatísticas oficiais sobre alunos inscritos e diplomados. Os resultados desta inquirição são ainda, após tratamento estatístico, transmitidos às instâncias internacionais EUROSTAT, OCDE e UNESCO, no âmbito dos compromissos internacionais assumidos na área das Estatísticas da Educação [23].

São objeto deste inquérito estatístico os cursos, conferentes ou não de grau académico, incluídos nos seguintes níveis de formação: Licenciatura – 1º ciclo, Mestrado Integrado, Mestrado – 2º ciclo, Doutoramento – 3º ciclo e Especialização pós-Licenciatura. Uma vez que são considerados cursos não conferentes de grau, é utilizada preferencialmente neste trabalho a designação de nível de formação, em vez de grau académico.

A DGEEC disponibiliza aos estabelecimentos de ensino a Plataforma de Recolha de Informação do Ensino Superior[3] (PRIES), através da qual deve ser submetida a resposta ao inquérito do RAIDES.

Para tal, os estabelecimentos dispõem de uma das seguintes opções de resposta: envio de um ficheiro no formato XML ou preenchimento de dados diretamente na plataforma eletrónica. A exportação de dados a enviar, ou o preenchimento *online*, relativo ao ano letivo N/N+1, é realizada em dois momentos diferentes:

- Momento 1: situação dos alunos em 31 de dezembro do ano N (decorre durante o mês de fevereiro);
- Momento 2: situação dos alunos em 31 de março do ano N+1 (decorre durante o mês de abril).

Neste trabalho foram usados os dados do momento 1, uma vez que os do momento 2 dizem respeito apenas aos alunos de doutoramento que ingressaram entre janeiro e março, e cujo número é residual.

No RAIDES são reportados cerca de 90 atributos (ou variáveis na nomenclatura utilizada no RAIDES) sobre o aluno, divididas nos seguintes quatro níveis:

- Identificação: 10 variáveis sobre a identificação do aluno;
- Inscrições: 39 variáveis que caraterizam a sua inscrição;
- Diplomas: 23 variáveis sobre o seu diploma;

[3] https://pries.dgeec.mec.pt.

- Mobilidade: 15 variáveis que caraterizam um aluno em mobilidade internacional (*incoming*).

Neste trabalho foram utilizados os níveis Identificação, Inscrições e Diplomas e o seguinte subconjunto de variáveis em cada um dos referidos níveis, mostrados nas tabelas com cabeçalho azul da Figura 3.7:

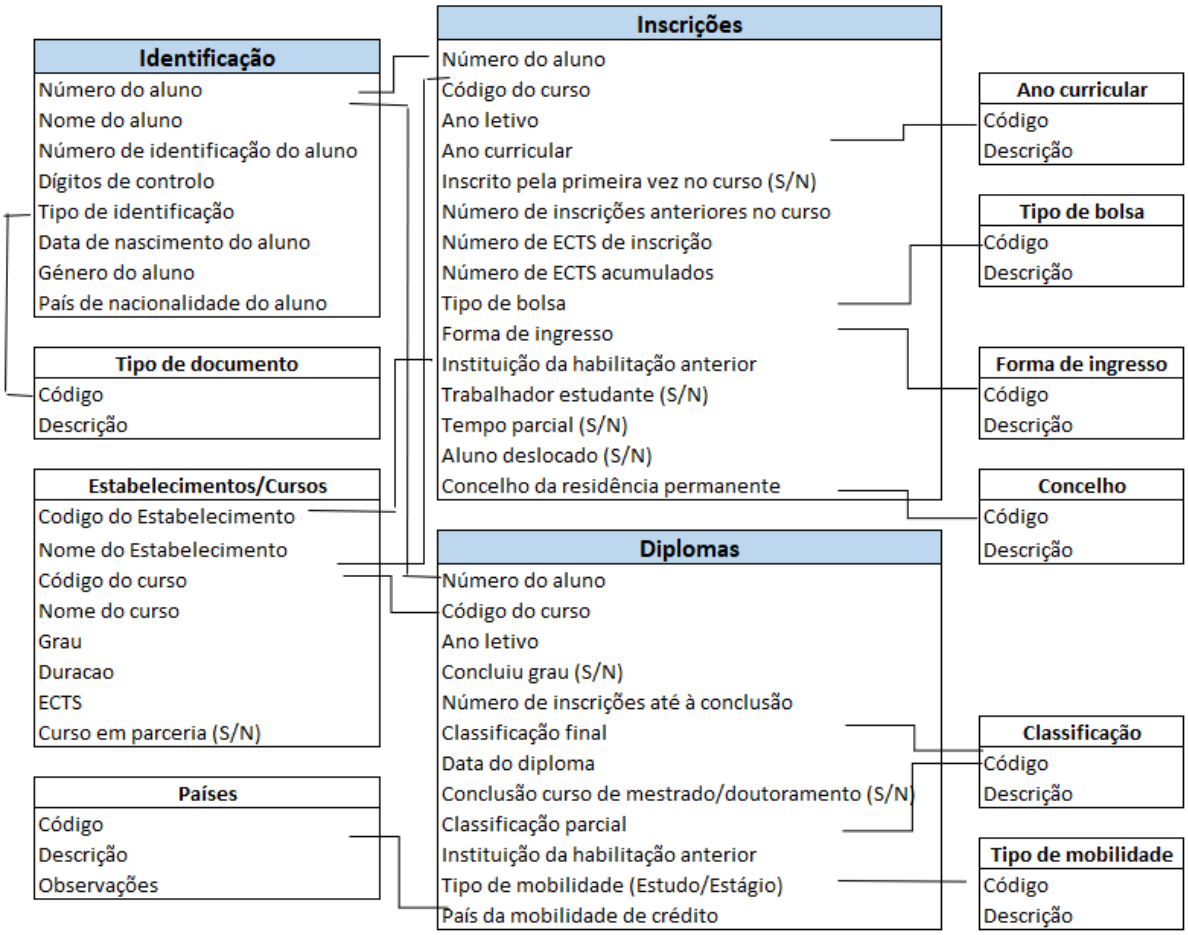


Figura 3.7 - Variáveis do RAIDES utilizadas neste trabalho e respetivas tabelas de suporte

No RAIDES existem 21 tabelas de suporte, necessárias às diferentes variáveis solicitadas na inquirição e que contêm os códigos utilizados no inquérito. As que foram utilizadas neste trabalho constam na Figura 3.7 com cabeçalho branco. A Tabela 3.3 apresenta uma descrição mais detalhada das tabelas de suporte e respetivas variáveis utilizadas neste trabalho. A coluna do Tamanho diz respeito ao número de dígitos, no caso dos campos de tipo numérico, e ao número de letras, no caso dos campos de tipo texto.

Tabela 3.3 - Tabelas de suporte do RAIDES e respetivas variáveis

Tabela	Variável	Tipo	Tamanho	Nº de linhas (em 2017)
TipoDocumento	Tipo do documento de identificação	Número	1	7
Países	País de nacionalidade	Texto	2	254
	País de mobilidade *outgoing*	Texto	2	254
Concelho	Concelho de residência permanente	Número	4	309

FormaIngresso	Forma de ingresso	Número	2	21
AnoCurricular	Ano curricular	Número	2	11
Bolseiro	Tipo de bolsa	Número	2	7
Classificação	Classificação final	Número	2	20
	Classificação final do curso de mestrado ou de doutoramento	Número	2	20
TipoMobilidade	Tipo de mobilidade *outgoing*	Número	1	2
Ficheiro MatrizCesRamos	Instituição de Ensino Superior anterior	Texto	4	40127
	Curso	Texto	4	40127

A DGEEC disponibiliza o ficheiro denominado "Matriz de estabelecimentos, cursos e ramos", no qual se encontram os cursos criados, os respetivos códigos e os ramos. Esta matriz contém ainda a informação relativa à duração, ECTS (*European Credit Transfer System*) e tipologia do curso (se é em parceria). Com a exceção da tabela dos cursos, que tem 18 colunas, todas as tabelas de suporte anteriores possuem apenas duas colunas: código e descrição (a dos Países contém adicionalmente uma coluna de observações).

Ao longo dos cinco anos considerados neste trabalho, a distribuição do número de alunos, inscrições e diplomas foi a seguinte:

Tabela 3.4 - Dados provenientes do RAIDES relativos aos últimos cinco anos letivos

Ano letivo	2013/14	2014/15	2015/16	2016/17	2017/18
Nº de Alunos	5913	5797	5937	5931	6053
Nº de Inscrições	5236	5171	5159	5183	5263
Nº de Diplomas	1239	1201	1345	1360	1421

As alterações que ocorreram durante estes cinco anos, nas tabelas do RAIDES, foram as seguintes:

- RAIDES17: Introdução de uma nova Forma de Ingresso na respetiva tabela;
- RAIDES16: Existência de um novo atributo no nível Identificação do Aluno: número de aluno (número associado ao aluno no estabelecimento de ensino). Na variável Forma de Ingresso, houve atualização das opções de ingresso, dando cumprimento à nova legislação, Portaria n.º 118/2015 de 19 de junho, que aprova o regulamento dos regimes de reingresso e de mudança de par instituição/curso no ensino superior [24];
- RAIDES15: Existência de um novo atributo nos diplomados de doutoramento, denominado Área FOS *(Fields of Science and Technology)*. Trata-se da classificação do domínio científico e tecnológico a atribuir à atividade de investigação efetuada.

No Capítulo 4 descreve-se qual a metodologia utilizada para dar resposta às referidas alterações.

O processo de resposta ao inquérito RAIDES decorre da seguinte forma: anualmente, durante o mês de janeiro, a DGEEC disponibiliza a informação relativa à calendarização do RAIDES, documentação de suporte à inquirição (ficheiro Excel com as diversas tabelas de suporte, ficheiro Excel dos

estabelecimentos e cursos, documento técnico da PRIES, documento de especificações da estrutura do ficheiro XML) e eventual informação sobre alterações previstas [23].

Simultaneamente, o GAAI inicia, na base de dados académica (Fenix), um processo de mapeamento de novos cursos/ramos ou ciclos de estudos alterados e de obtenção de relatórios que contêm informação sobre erros/avisos/informações, que vai sendo corrigida e/ou introduzida no próprio Fenix pela Direção Académica. Finalizadas as correções inicia-se um processo iterativo de submissão do ficheiro XML na PRIES e correção da informação, conforme apresentado na Figura 3.8:

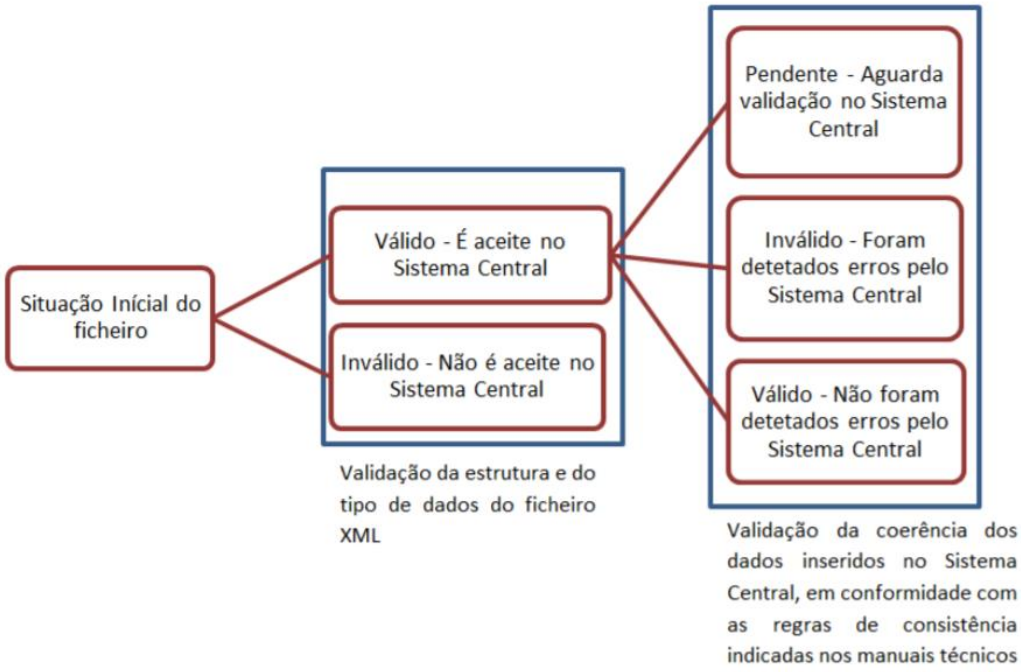


Figura 3.8 - Esquema da validação do ficheiro XML na PRIES [22]

Caso a estrutura e o tipo do ficheiro XML não sejam válidos, é reportada uma listagem de erros, que usualmente não ultrapassam a dezena, e que devem ser corrigidos para que o ficheiro seja aceite no sistema central. Uma vez que o ficheiro XML é aceite pelo sistema central, a validação da coerência dos dados pelo próprio sistema demora algumas horas/dias. Nas submissões iniciais os erros são na ordem das dezenas ou centenas. A título de exemplo: "o ano da habilitação anterior não está de acordo com o grau escolhido", "o documento de identificação não é válido" ou "o valor em bolseiro não está de acordo com o grau do curso".

Simultaneamente à deteção dos erros, a DGEEC disponibiliza uma validação de dados adicional que permite melhorar a qualidade dos dados reportados. A título de exemplo, algumas validações reportadas são: "inscritos em mais do que um curso", "reporte do género possivelmente incorreto", "datas de nascimento diferentes entre RAIDES17 e RAIDES16".

Existem, contudo, situações pontuais de erros (uma média de três situações anualmente) que acabam por ser corrigidas manualmente na própria PRIES: ou por motivo de cumprimento de prazos, uma vez que a PRIES demora horas/dias a validar e, caso apareçam novos erros a poucos dias de fechar a plataforma já não é possível voltar a submeter o ficheiro XML, ou por situações excecionais de última hora.

Uma vez realizada a confirmação dos dados submetidos, é encerrado o processo de resposta ficando bloqueada a PRIES, e a DGEEC considera o reporte de dados como finalizado. A partir desse momento

fica disponível um ficheiro Microsoft Excel, para *download*, que contém todos os dados reportados pelo estabelecimento, tal como estão guardados no Sistema Central da DGEEC [22].

O ficheiro Excel contém os dados nominais reportados, distribuídos por 4 folhas: Alunos, Inscrições, Diplomas e Mobilidade (*incoming*):

- Alunos: contém os elementos necessários para a identificação do aluno. Existe apenas uma identificação por aluno, ainda que este se possa encontrar em duas situações (exemplo: duas inscrições ou uma inscrição e um diploma). Deste modo, a uma identificação do aluno poderão ser associadas diferentes combinações de situações;
- Inscrições: contém a(s) inscrição(ões) do aluno no estabelecimento de ensino;
- Diplomas: contém o(s) diploma(s) que o aluno obteve. Neste ponto é importante referir que, desde o ano letivo 2010/11, são também contabilizados os diplomas de especialização atribuídos pela conclusão de um curso de mestrado ou de um curso de doutoramento, isto é, das partes curriculares dos referidos graus.

Os dados relativos à mobilidade (*incoming*) não foram considerados neste trabalho.

3.4 Indicadores académicos

De acordo com os referenciais para os sistemas internos de garantia de qualidade das IES, formulados pela Agência de Avaliação e Acreditação do Ensino Superior (A3ES), nomeadamente no Referencial nº 11 - *Gestão de Informação* [25], é mencionado que:

"A instituição [deve estar] dotada de mecanismos que permitam garantir a recolha, análise e utilização dos resultados e de outra informação relevante para a gestão eficaz dos cursos e demais atividades". Para isso, "[deve contar] com sistemas de recolha de informação fiáveis para o levantamento de resultados e outros dados e indicadores relevantes, que incluem, nomeadamente: indicadores-chave de desempenho; o perfil da população estudantil; as taxas de progressão, sucesso e abandono dos estudantes; a satisfação dos estudantes com os seus cursos; a empregabilidade e percursos profissionais dos graduados".

Adicionalmente no Referencial nº12 - *Informação Pública* [25], é referido que:

"A instituição [deve estar] dotada de mecanismos que permitam a publicação de informação clara, precisa, objetiva, atualizada, imparcial e facilmente acessível, acerca das atividades que desenvolve, (…) [incluindo] nomeadamente: a oferta formativa; os resultados do ensino, expressos nos resultados académicos, de inserção laboral e de grau de satisfação das partes interessadas (…)."

Num estudo adicional da A3ES sobre indicadores de desempenho para apoiar os processos de avaliação e acreditação de cursos [26] é mencionada a importância do conceito do valor acrescentado, referindo que "a avaliação das instituições deve levar em consideração as caraterísticas da população estudantil à entrada, de forma a medir, o desempenho da instituição e não só os resultados". Assim, a autora considera a existência de dois grupos de indicadores sobre o ensino: 1) caraterísticas dos estudantes e 2) desempenho dos estudantes.

Com base nos instrumentos de gestão interna, nas necessidades de reporte institucional, nos referenciais da A3ES e no referido estudo sobre indicadores para acreditação de cursos, foram identificados para este trabalho um conjunto de indicadores académicos, agrupados em: caraterização dos alunos, acesso, inscrição e conclusão. Saliente-se, contudo, que qualquer conjunto considerado nunca seria suficientemente exaustivo para cobrir todas as necessidades.

3.4.1 Caraterização dos alunos

Os indicadores sobre a caraterização dos alunos (CA) são transversais aos três processos de negócio e são os seguintes:

- **Número total de alunos num determinado ano letivo** (CA1): permite conhecer a dimensão, em termos de número de alunos, da FCUL;

- **Distribuição dos alunos segundo o género** (CA2): através deste indicador é possível fazer a caraterização da população estudantil quanto ao género;

- **Distribuição dos alunos segundo a idade** (CA3): permite caraterizar os alunos quanto à idade. Em relação a esta variável foi obtida a idade mínima, máxima, média e mediana, esta última por ser uma medida mais robusta;

- **Distribuição dos alunos de acordo com a origem geográfica** (CA4): permite avaliar a diversidade dos alunos relativamente à sua nacionalidade e avaliar os distritos de maior proveniência.

3.4.2 Acesso

Em relação à admissão de alunos, os seguintes indicadores de acesso avaliam a capacidade da FCUL para atrair novos alunos de licenciatura e mestrado integrado e caraterizam a qualidade dos alunos à entrada:

- **Índice de satisfação da procura** ou **número de candidatos em 1ª opção por vaga** (A1): quociente entre o número de candidatos em 1ª opção e o número de vagas, num determinado ano letivo (Equação (3.1)):

$$\text{A1} = \frac{\text{Número de candidatos em 1ª opção}}{\text{Número de vagas}} \quad (3.1)$$

- **Percentagem de alunos colocados em 1ª opção** (A2): quociente entre o número de alunos colocados em 1ª opção e o número de alunos colocados, num dado ano letivo (Equação (3.2)):

$$\text{A2} = \frac{\text{Número de colocados em 1ª opção}}{\text{Número de colocados}} \times 100 \quad (3.2)$$

- **Percentagem de ocupação de vagas** (A3): quociente entre o número de alunos colocados e o número de vagas, num determinado ano letivo (Equação (3.3)):

$$\text{A3} = \frac{\text{Número de colocados}}{\text{Número de vagas}} \times 100 \quad (3.3)$$

- **Média da nota de ingresso** (A4): quociente entre a soma da nota de candidatura dos alunos colocados e o número total de alunos colocados, num determinado ano letivo (Equação (3.4)):

$$\text{A4} = \frac{\sum(\text{Nota de candidatura})}{\text{Número de colocados}} \quad (3.4)$$

  Em relação à nota de ingresso, também foram obtidas as notas mínima e máxima dos alunos colocados.

3.4.3 Inscrição

Em relação aos alunos inscritos na FCUL, foram considerados os seguintes indicadores:

- **Percentagem de novos alunos** (I1): quociente entre o número de novos alunos (inscritos pela primeira vez no par instituição/curso) e o número total de alunos inscritos na FCUL, num dado ano letivo (Equação (3.5)):

$$\text{I1} = \frac{\text{Número de novos alunos}}{\text{Número total de inscritos}} \times 100 \quad (3.5)$$

- **Percentagem de alunos bolseiros** (I2): este indicador avalia as ajudas financeiras que são dadas aos alunos e se a instituição é acessível a todos os que pretendem frequentá-la. É obtido através do quociente, num dado ano letivo, entre o número de alunos que têm bolsa e o número total de alunos (Equação (3.6)):

$$\text{I2} = \frac{\text{Número de alunos bolseiros}}{\text{Número total de inscritos}} \times 100 \quad (3.6)$$

- **Número médio de ECTS inscritos** (I3): quociente entre a soma do número de ECTS em que os alunos se inscreveram, num determinado ano letivo, e o número total de alunos inscritos (Equação (3.7)):

$$\text{I3} = \frac{\sum(\text{Número de ECTS de inscrição})}{\text{Número total de inscritos}} \quad (3.7)$$

- **Percentagem de alunos inscritos exclusivamente em Estágio/Dissertação/Projeto** (I4): este indicador avalia a percentagem de alunos que estão inscritos exclusivamente em Estágio, Dissertação ou Projeto e por esse motivo não são contabilizados para financiamento pelo Ministério. É obtido através do quociente, num dado ano letivo, entre o número de alunos inscritos exclusivamente em Estágio/Dissertação/Trabalho de Projeto (EDT) e o número total de alunos de mestrado (Equação (3.8)):

$$\text{I4} = \frac{\sum(\text{Número de de alunos inscritos exclusivamente em EDT})}{\text{Número total de inscritos em mestrado}} \quad (3.8)$$

3.4.4 Conclusão

Em relação aos alunos diplomados na FCUL, ou seja, que concluíram os cursos, foram considerados os seguintes indicadores:

- **Classificação média dos diplomados** (D1): quociente entre a soma da classificação final dos alunos diplomados e o número total de alunos diplomados, num determinado ano letivo. Não é aplicável aos alunos de doutoramento uma vez que a classificação final nestes casos é qualitativa (Equação (3.9)):

$$\text{D1} = \frac{\sum(\text{Classificação final})}{\text{Número de diplomados}} \quad (3.9)$$

  No caso dos mestrados foi obtida adicionalmente a classificação média do curso de mestrado (parte curricular do mestrado).

- **Número médio de inscrições até à conclusão do curso** (D2): quociente entre a soma do número de inscrições dos alunos diplomados e o número total de alunos diplomados, num determinado ano letivo. Também foi calculada a mediana do número de inscrições até à conclusão (Equação (3.10)):

$$\text{D2} = \frac{\sum(\text{Número de inscrições})}{\text{Número de diplomados}} \quad (3.10)$$

- **Percentagem de diplomados com habilitação anterior obtida na FCUL** (D3): quociente entre o número de diplomados cujo grau académico anterior foi obtido na FCUL e o número de

diplomados, num determinado ano letivo. Não é aplicável a diplomados de licenciatura ou mestrado integrado (Equação (3.11)):

$$D3 = \frac{\text{Número de diplomados com habilitação anterior obtida na FCUL}}{\text{Número de diplomados}} \times 100 \quad (3.11)$$

- **Percentagem de diplomados que realizaram mobilidade (*outgoing*)** (D4): quociente entre o número de diplomados que realizaram mobilidade no seu percurso académico e o número de diplomados, num determinado ano letivo (Equação (3.12)):

$$D4 = \frac{\text{Número de diplomados que realizaram mobilidade}}{\text{Número de diplomados}} \times 100 \quad (3.12)$$

Qualquer um dos indicadores anteriores pode ser obtido, por nível de ensino (grau) ou por curso, dado que é informação relevante quer para a direção da FCUL, quer para os presidentes dos departamentos e coordenadores dos cursos.

Para além dos 16 indicadores identificados, foi também tida em conta a evolução de grande parte dos indicadores, no período em análise.

3.5 Fluxo de trabalho anterior

Atualmente no Portal de Ciências existe uma página de Estatísticas que engloba três vertentes principais, no que diz respeito a indicadores académicos, conforme a Figura 3.9:

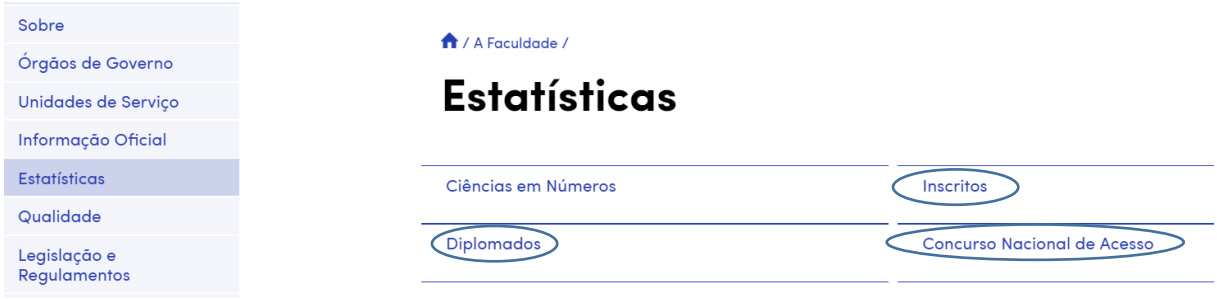


Figura 3.9 - Página de Estatísticas do Portal de Ciências

Cada uma destas três vertentes, que correspondem aos três processos de negócio académicos identificados na Secção 3.2, apresenta um ficheiro Excel por ano letivo (neste momento são apresentados os últimos seis anos letivos) e uma representação gráfica de cada uma das variáveis analisadas no referido ficheiro Excel. A Figura 3.10 apresenta um excerto da representação gráfica disponibilizada sobre os alunos inscritos:

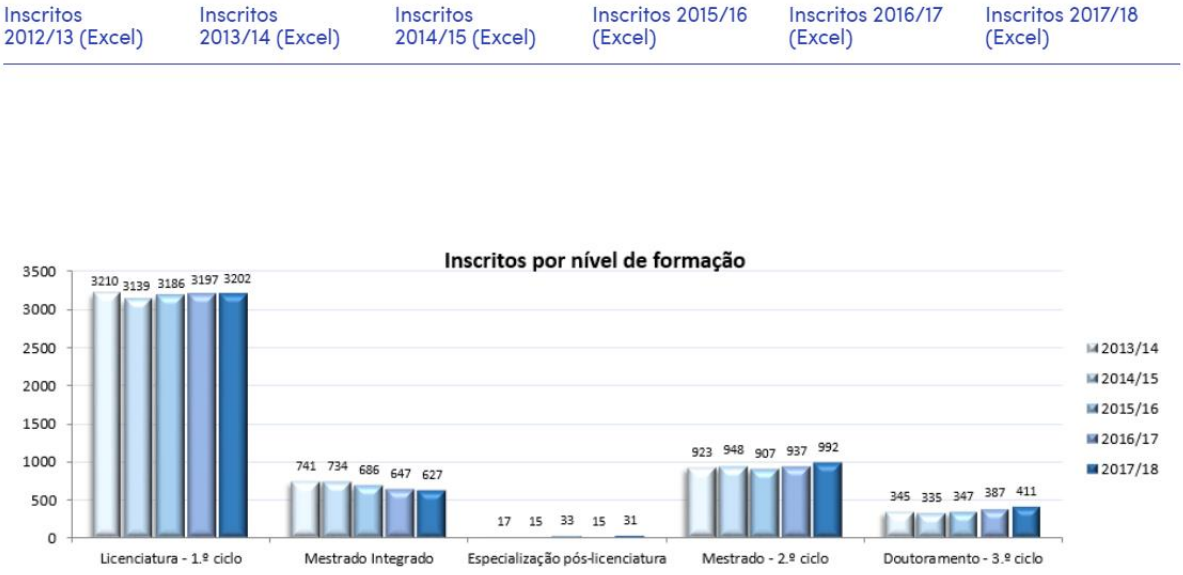


Figura 3.10 - Excerto da representação gráfica sobre inscritos, disponível no Portal de Ciências

Cada ficheiro Excel contém tabelas de frequências calculadas para diferentes variáveis, que correspondem, na sua grande maioria, aos indicadores sobre Caraterização dos alunos, identificados na Subsecção 3.4.1. A título de exemplo, no caso dos alunos inscritos e diplomados as variáveis existentes no ficheiro Excel são as que constam na Figura 3.11:

Figura 3.11 - Variáveis analisadas nos ficheiros sobre inscritos e diplomados no Portal de Ciências

A obtenção dos referidos ficheiros Excel e dos gráficos era realizada de forma manual e repetitiva, com os seguintes constrangimentos:

- Elevado tempo despendido na obtenção de indicadores (de uma a duas semanas);
- Impossibilidade de visualizar tendências/evoluções, uma vez que os ficheiros Excel apenas apresentam um determinado ano letivo;
- Problemas de desempenho no Excel, ao tentar integrar dados de vários anos letivos;

A Figura 3.12 exemplifica, para os dados sobre Inscritos e Diplomados, quais os procedimentos que eram realizados em termos de atualização anual, no fim de cada submissão do RAIDES:

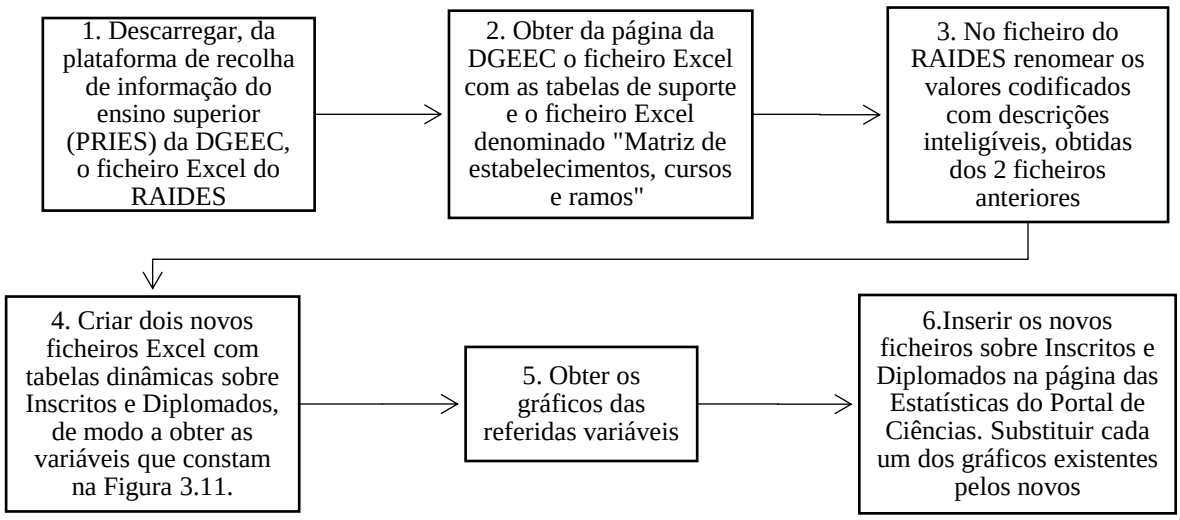


Figura 3.12 - Procedimento utilizado na análise de inscritos e diplomados

Durante o fluxo de trabalho anterior, as principais dificuldades estavam relacionadas com os seguintes pontos:

- Complexidade na obtenção de algumas métricas, nomeadamente o cálculo da mediana da idade e da mediana do número de inscrições, por curso;
- Necessidade de substituição de cada ficheiro e de cada gráfico, em caso de engano na fórmula ou na tabela;
- Dificuldade na obtenção dos diplomados totais e parciais em simultâneo numa única tabela dinâmica;
- Impossibilidade de ordenação das tabelas dinâmicas de agregados, pelo grau do curso e nome do curso, mas considerando também na visualização o atributo do código.

Assim, torna-se essencial que este processo realizado por um colaborador do GAAI, decorra de forma mais célere e eficiente, de modo a libertar recursos para outras análises de apoio ao planeamento e à decisão.

Torna-se também crucial que este processo deixe de ser realizado de forma manual e passe a ser automatizado, de cada vez que existem novos dados. Por outro lado, quando o número de dados no Excel começa a aumentar, o desempenho é afetado, pelo que é necessário um melhor desempenho e melhor capacidade analítica, com mais e melhores métricas.

3.6 Sumário

Este capítulo pretendeu dar uma visão mais específica do propósito do projeto, definindo os requisitos gerais através do levantamento de necessidades das diferentes partes interessadas, identificando os processos de negócio e respetivos indicadores académicos, apresentando as fontes de dados utilizadas no trabalho e descrevendo o fluxo de trabalho que era realizado no GAAI. Depois desta análise do problema, no capítulo seguinte vem a descrição da solução concretizada.

4. Concretização da solução

Depois de apresentados, nos capítulos anteriores, os conceitos teóricos relevantes num sistema de *Business Intelligence* (BI), os processos de negócio académicos considerados neste trabalho e as respetivas fontes de dados, o presente capítulo apresenta o desenho lógico do modelo dimensional da plataforma de indicadores académicos, com as tabelas de dimensões e factos, sendo depois descrito o processo de extração e tratamento dos dados dos alunos da Faculdade de Ciências (FCUL), para posterior carregamento num repositório único de modo a poderem ser utilizados na obtenção dos indicadores académicos e na construção dos relatórios interativos. No fim do capítulo é feita uma avaliação dos requisitos gerais definidos na Secção 3.1.

4.1 Visão geral

O desenvolvimento da solução de BI deste projeto seguiu as seguintes etapas recomendadas para o Power BI [27]:

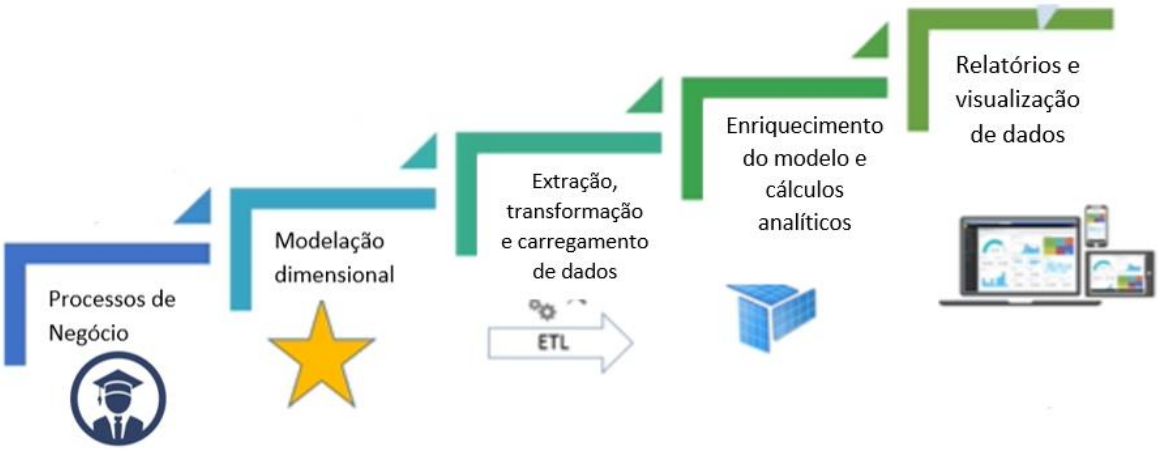


Figura 4.1 - Etapas de desenvolvimento da solução de BI neste trabalho [Adaptado de 27]

Assim, identificados os processos de negócio, seguiu-se o desenho do modelo dimensional. Este modelo deve ser simples e conseguir responder às perguntas de negócio. Depois, decorreu o processo de ETL (*Extract, Transform and Load*) para transformar a informação proveniente das diferentes fontes de dados. Uma vez o modelo consolidado, o passo seguinte foi melhorá-lo e enriquece-lo através da gestão das relações entre tabelas (de acordo com o modelo definido), criação de hierarquias, ordenação e agrupamento de dados e obtenção de algumas medidas e colunas calculadas, através da linguagem DAX *(Data Analysis eXpressions)*. Por último, foram elaborados os relatórios interativos, seguindo as boas práticas de construção de relatórios e *dashboards* [15].

A Figura 4.2 sintetiza os módulos da ferramenta Power BI Desktop utilizados em cada uma das etapas anteriores.

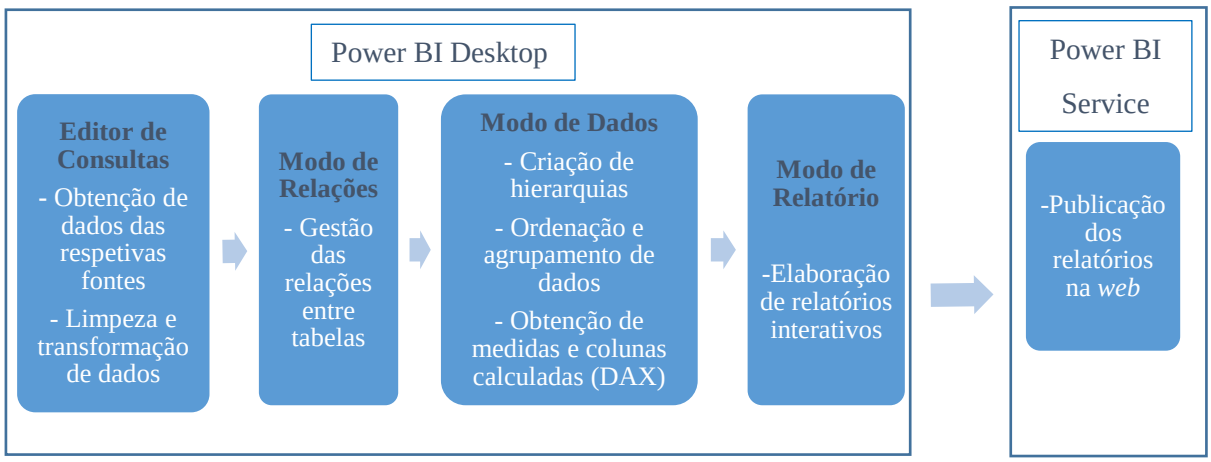


Figura 4.2 - Módulos utilizados no Power BI durante o fluxo de trabalho

A componente de ETL foi realizada no Editor de Consultas, apesar de uma parte significativa do processo de limpeza dos dados, também ter sido realizada no Excel. A gestão das relações entre tabelas foi feita no Modo de Relações, enquanto que a parte de criação de hierarquias, ordenação e agrupamento de dados, e obtenção de medidas calculadas e colunas calculadas (DAX) foi realizada no Modo de Dados. Por último, para a criação dos relatórios interativos foi utilizado o Modo de Relatório e, para a publicação dos dados, foi utilizado o Power BI Service.

4.2 Modelação dimensional

Na Secção 2.4 foram identificados os quatro passos recomendados na literatura para a modelação dimensional. O primeiro passo, relacionado com a definição dos processos de negócio prioritários, através da matriz de exequibilidade/valor, foi apresentado na Secção 3.2. Os três processos identificados para este trabalho foram os seguintes: acesso, inscrição e conclusão, dos alunos da FCUL. Prossegue-se agora com a identificação do grão das tabelas de factos, e com a descrição das tabelas de dimensões e das medidas.

4.2.1 Granularidade

Escolhidos os processos de negócio a modelar, é necessário determinar o grão das tabelas de factos. Cada processo de negócio contém pelo menos uma tabela de factos e respetivas dimensões, como indicado a seguir:

- **Acesso:** pretende-se a existência de dois grãos diferentes:
    1. Alunos colocados, numa determinada opção, num curso de Ciências, num ano letivo, através de um contingente, com uma determinada nota de candidatura;
    2. Total de alunos candidatos por curso, num determinado ano letivo;

  Um aluno que pretenda ingressar no ensino superior público, através do Concurso Nacional de Acesso (CNA), pode candidatar-se até um limite de seis pares instituição/curso. No final do processo de candidatura e após a seriação, o aluno passa de aluno candidato a aluno colocado num determinado par instituição/curso ou a aluno não colocado, caso não cumpra as condições necessárias em nenhuma das seis opções pretendidas.
- **Inscrição:** um aluno inscrito num curso de Ciências, num ano letivo, num determinado ano curricular, com um número de inscrições, de ECTS inscritos e de ECTS acumulados;
- **Conclusão:** um aluno diplomado (final ou parcial) num curso de Ciências, num determinado ano letivo, com uma classificação final e num determinado número de anos.

Nas secções seguintes, as quatro tabelas de factos foram denominadas da seguinte forma: Colocados, Agregados, Inscrições e Diplomas.

4.2.2 Dimensões

As tabelas de dimensões integram um conjunto diversificado de atributos, pelos quais os indicadores de negócio considerados nas tabelas de factos podem ser analisados.

A matriz de processos (*bus matrix)*, representada na Tabela 4.1, faz a associação entre os processos de negócio (linhas da matriz) e as dimensões definidas neste trabalho (colunas da matriz):

Tabela 4.1 - Matriz de processos

Processos de Negócio	Dimensões									
	Aluno	Aluno Colocado	Contingente	Curso	Data	Instituição	País	Perfil de Inscrição	Perfil do Aluno	Perfil do Diploma
Acesso		x	x	x	x					
Inscrição	x			x		x		x	x	
Conclusão	x			x	x	x	x			x

Verifica-se que a dimensão Curso é comum aos três processos de negócio. Já as dimensões Aluno e Instituição são comuns aos processos de inscrição e conclusão e a dimensão Data comum ao acesso e conclusão. As restantes dimensões são específicas de um só processo.

As três últimas dimensões apresentadas na Tabela 4.1 são minidimensões. Os seus atributos são textuais e não encaixam nas outras dimensões, pelo que foram agrupados em minidimensões.

De seguida são descritas as dez dimensões criadas neste trabalho. Em cada dimensão, para além da descrição dos atributos com conteúdo mais específico e da identificação da chave primária, são mencionadas as hierarquias consideradas.

Nas dimensões Aluno, Contingente, Curso e Instituição, apesar de não ser a regra habitual nos *data warehouses*, foi considerada a chave natural como chave primária uma vez que os códigos das chaves naturais são, nestes casos, atribuídos por Entidades oficiais que garantem a sua estabilidade e manutenção ao longo do tempo, não se prevendo qualquer alteração a este nível a médio/longo prazo. No caso do Aluno esta entidade é a própria FCUL. No caso do Contingente, é a Direção-Geral do Ensino Superior (DGES) e, no caso dos códigos do Curso e da Instituição de Ensino Superior é a Direção-Geral das Estatísticas de Educação e Ciência (DGEEC).

No caso das três minidimensões, as respetivas chaves substitutas foram criadas com base numa fórmula que combina os vários atributos envolvidos em cada uma delas.

Existe ainda a dimensão Opção de Candidatura, que corresponde a uma das seis opções em que o aluno é colocado no par instituição/curso. Trata-se de uma dimensão degenerada, isto é não tem mais nenhum atributo próprio, pelo que não tem uma tabela de dimensão associada, ficando diretamente na tabela de factos.

Uma vez que um *data warehouse* é um repositório de leitura de dados, e que a operação de escrita de dados está restringida ao carregamento e atualização do *data warehouse*, não é possível ao utilizador realizar qualquer alteração aos valores armazenados nas diversas dimensões. Mas, se for detetado um valor errado ou existir alguma alteração num determinado atributo, é necessário definir previamente qual a estratégia adotada para realizar tais atualizações, conforme mencionado na Secção 2.4. Todas

estas situações estão associadas a alterações que são pontuais, isto é, não muito frequentes e por isso as dimensões são designadas de dimensões de mudanças lenta (SCD – *Slowly Changing Dimensions*) [28]. Em algumas das dimensões a seguir descritas, é referido o tipo de mudança que foi considerada.

4.2.2.1 Dimensão Aluno

Esta dimensão guarda a informação referente às caraterísticas do aluno inscrito ou diplomado de Ciências:

Tabela 4.2 - Dimensão Aluno

Atributo	Tipo	Exemplo
Número do aluno (PK)	Número inteiro	28750
Nome do aluno	Texto	João Rodrigues da Silva
Número de identificação	Texto	YZ509677
Dígitos de controlo	Texto	2ZZ3
Tipo de identificação	Número inteiro	2
Descrição do tipo de identificação	Texto	Passaporte
Data de nascimento	Data	22/11/1990
Género	Texto	Masculino
País de nacionalidade	Texto	Portugal

A informação referente à dimensão Aluno é proveniente do ficheiro final dos dados do RAIDES (folha Alunos) e, antes da sua importação para o modelo, foi feita uma análise de modo a evitar alunos repetidos. Os casos repetidos foram eliminados guardando apenas uma linha de cada aluno. Nos casos em que existia uma data de nascimento diferente para o mesmo aluno, ou um número de identificação diferente para o mesmo aluno, a regra utilizada foi a de considerar sempre os dados do RAIDES mais recente. Isto é, foi utilizada a mudança lenta do tipo 1, uma vez que apenas se pretende a informação mais correta e atualizada.

Tendo em conta a experiência de anos anteriores do RAIDES, é usual que um número residual de alunos (0,1%) insira, na altura da sua inscrição, a data de nascimento incorreta e que no ano seguinte ao validar os seus dados, detete esta incorreção. Caso não seja detetada pelo aluno, a DGEEC durante o preenchimento do RAIDES, disponibiliza um conjunto de validações onde constam diferenças com o RAIDES anterior.

4.2.2.2 Dimensão Aluno Colocado

Esta dimensão guarda a informação referente ao aluno colocado através do CNA, num curso de licenciatura ou mestrado integrado, da FCUL:

Tabela 4.3 - Dimensão Aluno Colocado

Atributo	Tipo	Exemplo
Número do aluno (PK)	Número inteiro	3217
Nome do aluno	Texto	Maria de Sousa Carvalho

Número de identificação	Número inteiro	12409677
Data de nascimento	Data	08/02/1999
Género	Texto	Feminino
País de nacionalidade	Texto	Portugal
Distrito de residência	Texto	Setúbal
Escola do aluno	Texto	Liceu Camões
Distrito da escola	Texto	Setúbal

Um aluno colocado através do CNA pode não chegar a inscrever-se na FCUL. Anualmente, cerca de 8% dos alunos colocados na 1ª fase do CNA não efetivam a sua matrícula e inscrição, pelo que não chegam a ser alunos inscritos da FCUL.

A informação da dimensão Aluno Colocado é proveniente dos dados do Concurso Nacional de Acesso (CNA). Nas 4066 colocações de 1ª fase nos últimos 5 anos letivos, existiram 34 alunos colocados na FCUL em dois anos letivos diferentes, pelo que, nestes casos, foram eliminados os alunos repetidos, considerando apenas os dados do ano letivo mais recente. Esta regra, também utilizada nos dados provenientes do RAIDES, será usada para os anos letivos seguintes.

Os atributos País de Nacionalidade e Distrito de Residência, conforme referido na Secção 3.3.1, surgiram a partir do ano 2014, sendo que para registos anteriores a esse momento foi necessário adotar a designação "Não disponível".

4.2.2.3 Dimensão Contingente

Esta dimensão guarda a informação sobre o tipo de contingente (geral, deficientes, emigrantes, Açores ou Madeira) de candidatura do aluno:

Tabela 4.4 - Dimensão Contingente

Atributo	Tipo	Exemplo
Código do contingente (PK)	Número inteiro	17
Descrição do contingente	Texto	Geral

Neste caso, apesar de ser uma tabela com apenas dois atributos, não se recorreu a uma minidimensão, uma vez que, conforme referido anteriormente, são atributos da responsabilidade da DGES.

4.2.2.4 Dimensão Curso

Esta dimensão guarda a informação referente à oferta formativa da FCUL:

Tabela 4.5 - Dimensão Curso

Atributo	Tipo	Exemplo
Código oficial do curso (PK)	Texto	5572
Designação do curso	Texto	Engenharia Biomédica e Biofísica
Designação atual do curso	Texto	Engenharia Biomédica e Biofísica

Duração do curso (anos)	Número inteiro	3
Nível de formação	Texto	Doutoramento - 3.º ciclo
Código da área de estudos	Número inteiro	520
Área de estudos (2º nível)	Texto	Engenharia e Técnicas Afins
Código da área CNAEF	Número inteiro	524
Área CNAEF (3º nível)	Texto	Tecnologia dos Processos Químicos
Tipologia	Texto	FCUL
Departamento responsável	Texto	Departamento de Química e Bioquímica
Sigla do Departamento responsável	Texto	DQB
Ano de avaliação do curso	Texto	2019/20
Ano de extinção do curso	Texto	Em vigor

Uma vez que não existe nenhuma fonte de dados interna ou externa com a informação completa sobre os cursos lecionados numa determinada Instituição de Ensino Superior (IES), foi necessário compilar a informação desta dimensão de diferentes fontes: a grande maioria dos atributos anteriores foram obtidos do ficheiro de Matriz de estabelecimentos, Cursos e ramos disponibilizado anualmente pela Direção-Geral das Estatísticas de Educação e Ciência (DGEEC) na altura do preenchimento do Registo de Alunos Inscritos e Diplomados no Ensino Superior (RAIDES) [23].

Os códigos oficiais dos cursos, criados pela DGEEC, deixaram recentemente de ser numéricos e passaram a ser alfanuméricos. Contudo, esta alteração não teve implicações no desempenho da obtenção de relatórios deste trabalho: medições informais na escolha de diferentes filtros nas visualizações mostraram que todas elas se concretizaram em menos de cinco segundos. Por outro lado, em termos de carregamento de dados, tendo em conta a quantidade de dados utilizada neste trabalho, o desempenho também não foi afetado.

Foi criada uma coluna adicional denominada Designação atual do curso para dar resposta à atualização pontual na designação dos Cursos. Por exemplo, a licenciatura em Tecnologias de Informação e Comunicação alterou a designação para Tecnologias de Informação a partir do ano de 2015/16. No caso dos mestrados em Ensino, passaram de Mestrado em Ensino de Biologia e de Geologia a Mestrado em Ensino de Biologia e Geologia e de Mestrado em Ensino de Matemática no 3.º Ciclo do Ensino Básico e no Ensino Secundário a Mestrado em Ensino de Matemática no 3.º Ciclo do Ensino Básico e no Secundário.

No que diz respeito à Classificação Nacional de Áreas de Educação e Formação (CNAEF), a Portaria nº256/2016 de 16 de março estrutura as áreas em 3 níveis: grandes grupos (9 grupos), áreas de estudos (25 áreas) e áreas de educação e formação (77 áreas), habitualmente designadas por Áreas CNAEF. Neste projeto, os grandes grupos não foram considerados uma vez que a maioria dos cursos da FCUL pertencia apenas a um grupo (Ciências, Matemática e Informática).

A coluna Tipologia, fornece a informação sobre se o curso é exclusivamente da Faculdade de Ciências, "FCUL", se é "Inter-ULisboa", curso lecionado com outras Unidades Orgânicas (UO) da ULisboa ou se é um curso "em Associação", isto é, lecionado com outras IES.

Por último, foram adicionadas três colunas: a coluna do Departamento Responsável pelo curso, sendo que nos cursos lecionados com outras instituições e em que a FCUL não seja a entidade responsável pelo curso, foi atribuído o departamento "Externo"; outra coluna, denominada Ano de Avaliação do curso identifica, para o próximo quinquénio (2017/18 a 2021/22), qual o ano em que o curso será avaliado pela Agência de Avaliação e Acreditação do Ensino Superior (A3ES); e, finalmente, Ano de extinção do curso é um atributo preenchido com o ano de extinção para os cursos já extintos, ou com "Em vigor" para os atuais.

Conforme descrito na Secção 2.4, uma tabela de dimensão pode ser constituída por atributos que se distribuem em vários níveis hierárquicos. É possível que uma dimensão inclua mais do que uma hierarquia. Assim, nesta dimensão existem três hierarquias:

- Nível de formação > Designação atual do curso
- Área de Estudos > Área CNAEF
- Departamento responsável > Designação atual do curso

A dimensão Curso é atualizada aquando da criação de um novo ciclo de estudos, tendo existido uma média de dois cursos novos por ano.

4.2.2.5 Dimensão Data

Esta dimensão, muito habitual nos *data warehouses*, é essencial para analisar determinado processo ao longo do tempo. É das poucas dimensões que pode ser construída de antemão e, neste caso, foi criada desde 2012 a 2019.

Tabela 4.6 - Dimensão Data

Atributo	Tipo	Exemplo
Data (PK)	Data	20151012
Data	Data	12/10/2015
Dia	Número inteiro	12
Dia da semana	Texto	Segunda-feira
Número do mês	Número inteiro	10
Nome do mês	Texto	Outubro
Ano	Número inteiro	2015
Ano letivo	Texto	2015/16
Semestre letivo	Texto	1º semestre

No caso da dimensão Data, a chave substituta foi criada de acordo com as regras recomendadas para este tipo de dimensão "YYYYMMDD" [10].

A dimensão Data é hierárquica, uma vez que muitos dos seus atributos têm relações de hierarquia:

- Ano > Número do mês > Dia
- Ano letivo > Semestre letivo

4.2.2.6 Dimensão Instituição

Esta dimensão guarda a informação da instituição de ensino superior onde o aluno obteve a sua habilitação anterior (apenas para alunos de mestrado e doutoramento e desde que tenha sido obtida em Portugal):

Tabela 4.7 - Dimensão Instituição

Atributo	Tipo	Exemplo
Código da Instituição	Número inteiro	1100
Designação da Instituição	Texto	Universidade do Porto
Código da Unidade Orgânica (PK)	Número inteiro	1103
Designação da Unidade Orgânica	Texto	Universidade do Porto - Faculdade de Ciências
Tipo de Estabelecimento	Texto	Público
Tipo de Ensino	Texto	Universitário
NUTS II	Texto	Norte
Distrito	Texto	Porto
Concelho	Texto	Porto

As NUTS, Nomenclatura das Unidades Territoriais para Fins Estatísticos, são uma classificação territorial comum adotada em todos os países da União Europeia para fins estatísticos [29]. No caso das NUTS II, existem 7 unidades territoriais: Norte, Centro, Área Metropolitana de Lisboa, Alentejo, Algarve, Região Autónoma da Madeira e Região Autónoma dos Açores.

Nesta dimensão existem duas hierarquias:

- Designação da Instituição > Designação da Unidade Orgânica
- NUTS II > Distrito > Concelho

4.2.2.7 Dimensão País

Esta dimensão guarda a informação referente ao País de nacionalidade do aluno ou ao país em que o aluno realizou mobilidade *outgoing*:

Tabela 4.8 - Dimensão País

Atributo	Tipo	Exemplo
Código do País (PK)	Número inteiro	24
Sigla do País	Texto	PT
Designação do País	Texto	Portugal
Continente	Texto	Europa

No caso da dimensão País, a chave substituta foi criada através de um número sequencial.

Nesta dimensão existe a hierarquia Continente > Designação do País.

4.2.2.8 Dimensão Perfil de Inscrição

Esta minidimensão guarda a informação relativa à forma de ingresso, para os novos alunos de licenciatura ou mestrado integrado, caso contrário está preenchido com "Forma de ingresso não aplicável". Guarda também, neste caso para todos os alunos, a informação relativa ao ano curricular em que o aluno está inscrito:

Tabela 4.9 – Dimensão Perfil de Inscrição

Atributo	Tipo	Exemplo
Código do Perfil da inscrição (PK)	Número inteiro	1010
Descrição da forma de ingresso	Texto	Regime Geral de Acesso
Ano curricular	Texto	Dissertação

De acordo com as tabelas de códigos fornecidas pela DGEEC [22], a forma de ingresso pode assumir 21 valores diferentes. Estas categorias têm sido atualizadas por motivo de alteração da legislação mas tem havido o cuidado de as novas formas de ingresso serem adicionadas sequencialmente, mantendo sempre o histórico, o que não tem implicações nos anos anteriores.

No que diz respeito ao ano curricular, de acordo com as tabelas de códigos fornecida pela DGEEC, esta dimensão pode assumir os seguintes valores: 1.º, 2.º, 3.º, 4.º, 5.º, Estágio final, Trabalho de Projeto e Dissertação. As opções Estágio final, Trabalho de Projeto e Dissertação aplicam-se apenas a alunos de mestrado ou mestrado integrado que estejam exclusivamente inscritos numa destas três opções. No caso dos alunos de doutoramento, o ano curricular é preenchido com "Ano curricular não aplicável".

4.2.2.9 Dimensão Perfil do Aluno

Esta minidimensão guarda a informação relativa aos diferentes estatutos que o aluno pode ou não ter, nomeadamente bolseiro, trabalhador estudante, tempo parcial e deslocado da residência principal:

Tabela 4.10 - Dimensão Perfil do Aluno

Atributo	Tipo	Exemplo
Código do Perfil do aluno (PK)	Número inteiro	10015
Descrição do tipo de bolsa	Texto	Bolseiro da ação social do ensino superior
Trabalhador Estudante	Texto	Não Trabalhador-Estudante
Tempo parcial	Texto	Tempo integral
Deslocado	Texto	Aluno deslocado

De acordo com as tabelas de códigos fornecida pela DGEEC [22], o atributo Descrição do tipo de bolsa pode assumir sete valores distintos, sendo algumas opções apenas válidas para alunos de licenciatura, mestrado integrado ou mestrado (Bolseiro da ação social do ensino superior) e outras apenas para alunos de doutoramento (Bolseiro da Fundação para a Ciência e a Tecnologia).

No que diz respeito aos restantes atributos desta minidimensão, no ficheiro original do RAIDES assumem apenas os valores verdadeiro ou falso, mas na tabela de dimensão passam a ter as designações

completas: Trabalhador Estudante/Não Trabalhador-Estudante, Tempo parcial/Tempo integral, Aluno deslocado/Aluno não deslocado.

4.2.2.10 Dimensão Perfil do Diploma

Esta minidimensão guarda a informação referente ao tipo de diploma que o aluno obteve e ao tipo de mobilidade que o aluno realizou durante o seu percurso formativo, possuindo os seguintes atributos:

Tabela 4.11 - Dimensão Perfil do Diploma

Atributo	Tipo	Exemplo
Código do Perfil do diploma (PK)	Número inteiro	10
Descrição do diploma	Texto	Diploma final
Descrição da mobilidade	Texto	Mobilidade de estudos

O atributo Descrição do Diploma, pode assumir três valores: diploma final, diploma parcial, e diploma final e parcial. O diploma parcial apenas existe no caso dos mestrados e doutoramentos, e corresponde à conclusão da parte curricular dos referidos graus. O atributo da mobilidade pode assumir três valores: mobilidade de estudos, mobilidade de estágio e sem mobilidade.

4.2.3 Medidas

Conforme referido na Subsecção 4.2.1, relativa à granularidade, para além das três tabelas de factos em que o grão definido foi o aluno, existiu a necessidade, em relação aos dados dos alunos candidatos, de criar uma tabela de factos ao nível do curso, uma vez que a caraterização dos candidatos apenas precisa de ser feita de forma agregada.

Todas as tabelas de factos existentes neste trabalho são do tipo transacional, em que cada linha representa uma transação ou acontecimento. Ou seja, cada tabela de factos regista eventos que ocorrem em determinados momentos não periódicos, onde, para cada linha, haverá apenas uma data associada e não haverá atualização de factos existentes.

Na tabela de factos dos Colocados, a medida utilizada foi a seguinte:

- Nota de candidatura, a qual serviu para obter o indicador académico Média da Nota de Ingresso (A4).

No facto das Inscrições foram consideradas três medidas:

- Número de inscrições anteriores no curso;
- Número de ECTS acumulados (apenas para alunos com mais do que uma inscrição);
- Número de ECTS inscritos, o qual serviu para obter o indicador Número Médio de ECTS Inscritos (I3).

No caso dos factos dos Diplomas, foram identificadas duas medidas:

- Classificação final;
- Número de anos até à conclusão,

que serviram respetivamente para obter os indicadores Classificação Média dos Diplomados (D1) e Número Médio de Inscrições até à Conclusão do Curso (D2).

As medidas anteriores são não aditivas, ou seja, não podem ser somadas de acordo com qualquer uma das dimensões. Podem, no entanto, ser calculadas médias, mínimos, máximos, medianas.

4.2.4 Modelo de dados

Tendo em conta a granularidade, as dimensões e as medidas descritas anteriormente, o modelo final dos dados é o seguinte:

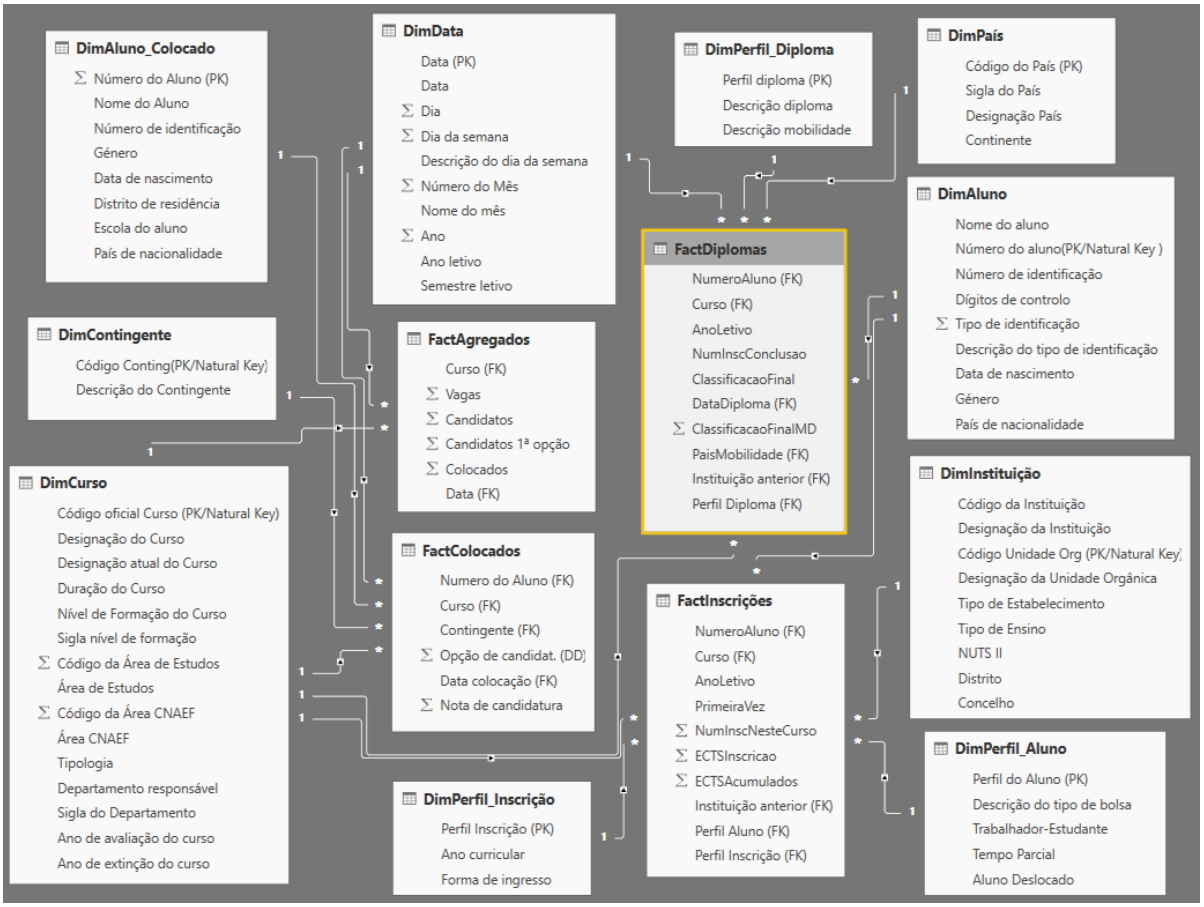


Figura 4.3 - Modelo multidimensional final dos dados

As tabelas de factos da Figura 4.3 contêm as colunas das chaves estrangeiras que ligam às diferentes tabelas de dimensões, e das medidas. A tabela de factos das Inscrições contém, para além das chaves estrangeiras e das medidas, uma *flag* para distinguir os alunos em que é a primeira inscrição (valor 1) dos alunos em que não é (valor 0).

4.3 Extração, transformação e carregamento de dados

Uma vez identificadas as fontes de dados e desenhado o modelo dimensional de dados da plataforma de indicadores académicos escolhidos neste trabalho, foi necessário trabalhar os dados, através de uma série de etapas do processo de ETL:

a) extração dos dados desde a sua origem;
b) transformações, deteção de incoerências e validações;
c) carregamento dos dados prontos a usar pelos decisores.

4.3.1 Extração de dados

Uma vez que este trabalho diz respeito a uma análise evolutiva de cinco anos, a primeira etapa foi juntar os dados relativos a este horizonte temporal, para cada uma das fontes de dados. Tendo em conta a existência de mudanças nas tabelas e atributos ao longo do tempo, descritas na Secção 3.3, a primeira ação foi verificar a existência de novas tabelas de suporte, de novos atributos nas tabelas de suporte ou de novos códigos ou descrições em atributos já existentes, de modo a que a integração fosse coerente.

4.3.1.1 Dados do Concurso Nacional de Acesso

No caso dos dados provenientes do Concurso Nacional de Acesso (CNA), em formato Access, foi utilizada a seguinte metodologia: foram importadas para o ficheiro Access que continha os dados de 2017, as tabelas Alunos, AlunosIdentidade, Alunos Cand e AlunosPrefer relativas aos anos compreendidos entre 2013 e 2016. Posteriormente, com os dados centralizados num único ficheiro Access, foram feitas as consultas necessárias.

Após a fusão da Universidade de Lisboa com a Universidade Técnica de Lisboa, os códigos oficiais das Unidades Orgânicas foram alterados, motivo pelo qual foi necessário utilizar um código no ano de 2013 (0701) e o novo código (1503) a partir do ano 2014.

4.3.1.2 Dados do RAIDES

No caso do RAIDES foi utilizada uma metodologia idêntica, mas neste caso juntando os cinco ficheiros Excel, mencionados na Subsecção 3.3.2 e referentes aos anos compreendidos entre 2013 e 2017, num único ficheiro Excel. Uma vez que ao longo destes cinco anos foi criado um novo atributo na tabela Aluno, o Número de aluno, e um novo atributo na tabela Diplomas, a área FOS, foi necessário uniformizar as tabelas existentes de modo a que as colunas com a mesma informação ficassem coerentes.

4.3.1.3 Repositório Excel e Editor de Consultas

Depois da integração dos dados provenientes de cada uma das fontes, todos os dados foram guardados num único ficheiro Excel (repositório). É importante existir um repositório único dos dados, caso contrário mantém-se o problema da existência de diversos ficheiros para os mesmos conjuntos de dados. Por outro lado, e uma vez que os dados uma vez carregados no Power BI Desktop não são exportáveis, seguiu-se esta metodologia de juntar todas as tabelas de factos e dimensões num único ficheiro Excel.

Uma vez guardados os dados num único repositório, foi necessário realizar várias transformações sobre os mesmos na ferramenta Power BI. No Power BI Desktop podem obter-se dados de uma variedade de fontes, incluindo bases de dados relacionais, ficheiros e serviços *online*. Depois de escolher a fonte de dados pretendida, neste caso o ficheiro Excel do repositório de dados com todas as tabelas de factos e dimensões, a primeira janela que aparece no Power BI é a do Navegador. O Navegador mostra as tabelas ou entidades da fonte de dados e, ao clicar nelas aparece a visualização do seu conteúdo. Depois de selecionar todas as folhas do ficheiro Excel, existem duas opções: 1) importar as tabelas imediatamente para o Power BI Desktop, através do botão Carregar ou 2) selecionar o botão Editar para transformar e limpar os dados antes da sua importação, com o Editor de Consultas (Power Query). A opção 2) é a mais conveniente ao trabalhar com grandes conjuntos de dados uma vez que, editar uma consulta antes de a carregar, permite reduzir o volume de dados antes do carregamento.

A Figura 4.4 apresenta a obtenção das 14 tabelas de dados existentes no ficheiro do repositório e o seu carregamento no Editor de Consultas do Power BI Desktop, para poderem ser trabalhadas e transformadas.

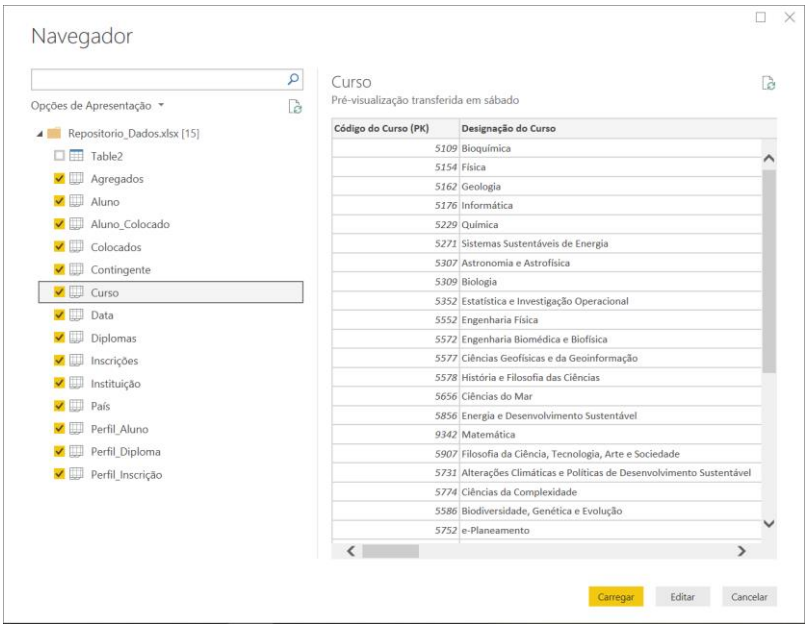


Figura 4.4 - Obtenção de todas as tabelas de dados utilizadas neste trabalho e carregamento no Editor de Consultas

Uma vez obtidas do repositório Excel todas as tabelas utilizadas no trabalho, seguiu-se a fase de transformação.

4.3.2 Transformação de dados

Os dados contidos no repositório Excel sofreram um tratamento, antes de serem carregados para o Power BI, que consistiu essencialmente na correção de erros, introdução de informação em falta e uniformização dos atributos existentes. Esta última correção de informação, com o objetivo de uniformizar, foi necessária devido, quer às mudanças de regras da DGEEC, quer à mudança da base de dados académica, no ano de 2016. Por exemplo, no sistema académico anterior, em funcionamento até ao ano letivo 2015/16, caso a informação constasse no sistema, era preenchida no ficheiro do RAIDES, mesmo que fosse de carater opcional. No sistema académico atual, essa informação apenas é preenchida e enviada no RAIDES para os atributos em que ela é obrigatória. Caso contrário é enviada a *null*.

4.3.2.1 No Repositório Excel

De seguida é descrito o tratamento inicial dos dados provenientes do RAIDES, sendo que nos dados do CNA não existiu a necessidade de nenhuma correção.

Correções de erros

Correção 1: Datas de diplomas

- No ficheiro Excel do RAIDES16, na folha Diplomas constava o valor 1900, na coluna da data do diploma, em três diplomados parciais que tinham sido inseridos manualmente na Plataforma de Recolha de Informação do Ensino Superior (PRIES);
- Linhas afetadas: 3;
- Resolução: Apagar os referidos valores, uma vez que essa coluna apenas deve ser preenchida para diplomados finais e não para diplomados parciais;

Correção 2: Número de aluno

- No ficheiro Excel do RAIDES16, na folha Alunos constava um aluno com um Número de aluno diferente ao RAIDES15;
- Linhas afetadas: 1;

- Resolução: Corrigir no RAIDES15 para o número mais atual (do RAIDES16);

Correção 3: ECTS acumulados

- No ficheiro Excel do RAIDES15, na folha Inscrições constavam alunos inscritos pela primeira vez, com a coluna ECTS acumulados preenchida;
- Linhas afetadas: 3;
- Resolução: Apagar os referidos valores uma vez que essa coluna apenas pode ser preenchida para alunos primeira vez=falso;

Introdução de informação em falta

Correção 4: Datas de conclusão

- No ficheiro Excel do RAIDES13 e do RAIDES14, na folha Diplomas estavam em falta as datas de conclusão dos alunos graduados das licenciaturas dos mestrados integrados;
- Linhas afetadas: 97 (RAIDES13) + 78 (RAIDES14);
- Resolução: Obtê-las do sistema académico e introduzi-las no ficheiro final;

Correção 5: Número de aluno

- No ficheiro Excel do RAIDES13, RAIDES14 e RAIDES15 na folha Alunos inserir a coluna com o Número de aluno, novo atributo introduzido a partir do RAIDES16;
- Linhas afetadas: 5913 (RAIDES13) + 5797 (RAIDES14) + 5937 (RAIDES15);
- Resolução: Obtê-las do sistema académico e introduzi-las no ficheiro;

Correção de informação com vista à uniformização

Correção 6: Atributo Primeira vez

- No ficheiro Excel do RAIDES13, na folha Inscrições constavam alunos com Número de inscrições anteriores=1 e primeira vez=verdadeiro;
- Linhas afetadas: 47;
- Resolução: Corrigir para primeira vez=falso;

Correção 7: Datas de conclusão

- No ficheiro Excel do RAIDES14 e do RAIDES15, na folha Diplomas constavam as datas de conclusão em alunos graduados parciais;
- Linhas afetadas: 291 (RAIDES14) + 292 (RAIDES15);
- Resolução: Apagar os referidos valores uma vez que essa coluna apenas é obrigatória para diplomados finais e, no RAIDES16 e RAIDES17, não sendo obrigatória, não foi preenchida para diplomados parciais;

Correção 8: Formas de ingresso

- No ficheiro Excel do RAIDES15, na folha Inscrições constavam formas de ingresso em alunos de mestrado e doutoramento;
- Linhas afetadas: 19;
- Resolução: Apagar os referidos valores uma vez que essa coluna apenas é obrigatória para alunos de licenciatura e mestrado integrado;

Correção 9: Número de inscrições

- No ficheiro Excel do RAIDES16, na folha Inscrições constavam alunos inscritos pela primeira vez, com número de inscrições anteriores preenchido com 0 anos;
- Linhas afetadas: 6;
- Resolução: Apagar os referidos valores uma vez que essa variável apenas é preenchida com valores maiores ou iguais a um;

Correção 10: ECTS acumulados

- No ficheiro Excel do RAIDES13, RAIDES14 e RAIDES15 na folha Inscrições constavam alunos inscritos pela primeira vez, com a coluna ECTS acumulados preenchida com 0;
- Linhas afetadas: 4357;
- Resolução: Apagar os referidos valores, uma vez que essa coluna apenas deve ser preenchida para primeira vez=falso;

Após esta limpeza inicial realizada no próprio ficheiro Excel do repositório com o objetivo de corrigir erros, completar informação e uniformizá-la, as transformações subsequentes foram realizadas no Power BI.

4.3.2.2 No Editor de Consultas do Power BI

O objetivo das transformações realizadas no Power BI foi tornar os dados mais compreensíveis e incluiu, entre outros: a eliminação ou renomeação de colunas, a substituição de valores para os tornar inteligíveis, a união/acrescento de consultas e a correção dos tipos de dados.

Cada coluna tem um determinado tipo de dados, que é atribuído automaticamente na importação dos dados para o Power BI, segundo o tipo de dados da coluna de origem. Contudo, este tipo de dados deve ser revisto e, caso necessário, corrigido. É importante referir que o tipo de dados (texto, número, data) afeta o armazenamento, enquanto o formato (por exemplo na data: dia, ou dia/mês) só afeta a visualização.

A ferramenta do Power BI para formatação e transformação de dados, de modo a que fiquem prontos para a modelação e visualização, é o Editor de Consultas, descrita na Subseção 2.7.2.

Na Figura 4.5, os Passos Aplicados, do painel Definições da Consulta, refletem os que foram aplicados na dimensão Aluno:

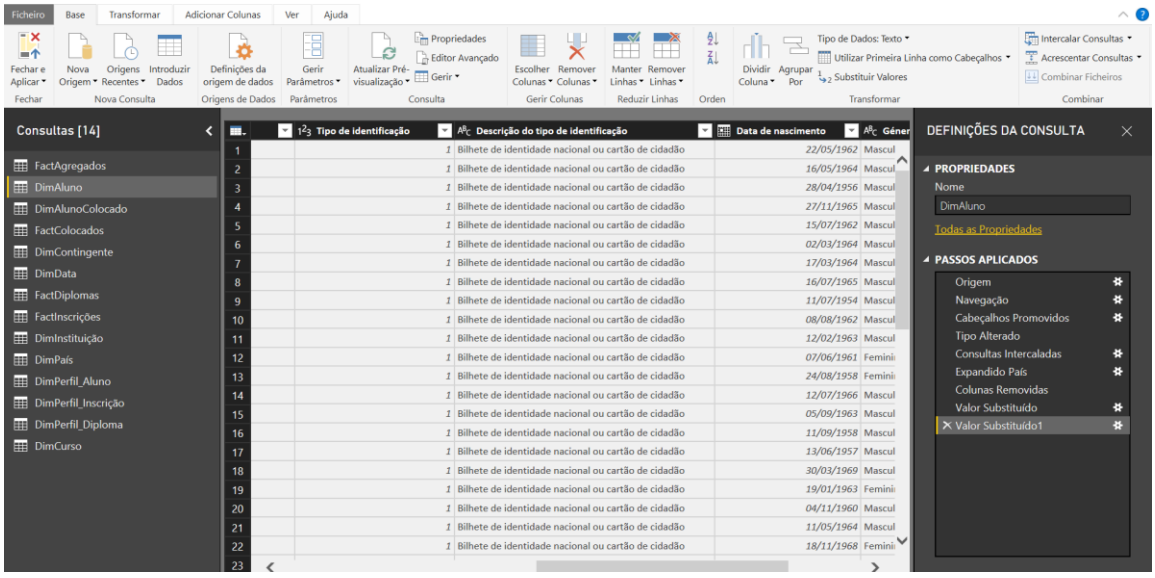


Figura 4.5 - Transformações aplicadas no Editor de Consultas, em relação à dimensão Aluno

a) Os quatro primeiros passos são executados pelo Power BI em qualquer consulta: ligar ao repositório dos dados (Origem), selecionar a tabela (Navegação), utilizar a 1ª linha como cabeçalho (Cabeçalhos Promovidos) e, ao carregar as tabelas, alterar automaticamente (de texto para número inteiro) as colunas com números em formato de texto (Tipo Alterado);
b) Os seguintes passos foram os necessários especificamente na dimensão Aluno e estão descritos de forma mais detalhada na Tabela 4.12.

Se uma transformação não funcionar da maneira pretendida ou se for necessário remover qualquer etapa, basta selecionar o X ao lado da etapa.

Todas as etapas da formatação ou transformação de uma consulta para além de serem registadas no painel Passos Aplicados do Editor de Consultas, também o são no Editor Avançado, gerando internamente uma série de expressões em linguagem M. A Figura 4.6 apresenta os passos aplicados que ficaram registados, em linguagem M, para a dimensão Aluno, no Editor Avançado do Power BI:

```
let
    Origem = Excel.Workbook(File.Contents("C:\Repositorio_Dados.xlsx"), null, true),
    Aluno_Sheet = Origem{[Item="Aluno",Kind="Sheet"]}[Data],
    #"Cabeçalhos Promovidos" = Table.PromoteHeaders(Aluno_Sheet, [PromoteAllScalars=true]),
    #"Tipo Alterado" = Table.TransformColumnTypes(#"Cabeçalhos Promovidos",{{"Nome do aluno", type text}, {"Número do aluno", Int64.Type}, {"Número de identificação", type any}, {"Dígitos de controlo", type text}, {"Tipo de identificação", Int64.Type}, {"Descrição do tipo de identificação", type text}, {"Data de nascimento", type date}, {"Género", type text}, {"País de nacionalidade", type text}}),
    #"Consultas Intercaladas" = Table.NestedJoin(#"Tipo Alterado",{"País de nacionalidade"},DimPaís,{"Sigla País"},"País",JoinKind.LeftOuter),
    #"Expandido País" = Table.ExpandTableColumn(#"Consultas Intercaladas", "País", {"Designação País"}, {"Designação País"}),
    #"Colunas Removidas" = Table.RemoveColumns(#"Expandido País",{"País de nacionalidade"}),
    #"Valor Substituído" = Table.ReplaceValue(#"Colunas Removidas","M","Feminino",Replacer.ReplaceText,{"Género"}),
    #"Valor Substituído1" = Table.ReplaceValue(#"Valor Substituído","H","Masculino",Replacer.ReplaceText,{"Género"})
in
    #"Valor Substituído1"
```

Figura 4.6 - Transformações registadas no Editor Avançado, em relação à dimensão Aluno

A Tabela 4.12 apresenta uma descrição detalhada de todos os passos aplicados no Editor de Consultas durante a etapa de transformação dos dados.

Tabela 4.12 - Passos aplicados neste trabalho durante a etapa de transformação

Propósito	Designação dos Passos aplicados	Descrição	Tabelas
(Executado por omissão)	Origem	Ligar ao repositório dos dados	Todas as tabelas
	Navegação	Selecionar a tabela	
	Cabeçalhos Promovidos	Considerar a 1ª linha como cabeçalho	
	Tipo Alterado	Alterar automaticamente (de texto para número inteiro) as colunas com números em formato de texto	

Obter a designação do País	Consultas Intercaladas Expandido País Colunas Removidas	Intercalar o atributo País de nacionalidade das dimensões Aluno e Aluno Colocado, com o atributo Designação do País da dimensão País	DimAluno, DimAlunoColocado
Inserir nomes inteligíveis nos valores dos atributos	Valor substituído	Atributo Género: M = Feminino e H = Masculino Atributo Género: M = Masculino e F = Feminino Atributo Primeira Vez: Verdadeiro = 1 e Falso = 0 Atributo Designação do País: *Null* = Não disponível	DimAluno DimAlunoColocado FactInscrições DimAlunoColocado
Remover colunas não utilizadas	Outras colunas removidas	Eliminar algumas colunas existentes no ficheiro Excel do RAIDES e não utilizadas neste trabalho	FactInscrições e FactDiplomas
Ligar tabelas	Consultas Intercaladas Expandido País/Data Colunas com Nome Mudado Colunas Removidas	Ligação das chaves substitutas das dimensões às chaves estrangeiras das tabelas de factos	FactColocados FactInscrições FactDiplomas
Verificar tipos de dados e inserir nomes inteligíveis nos atributos	Tipo Alterado Colunas com Nome Mudado	Verificar se os tipos de dados das colunas estão bem definidos e, em caso de necessidade, corrigi-los Corrigir designações de atributos para que fiquem inteligíveis	Todas as tabelas
Ordenar atributo	Personalizado adicionado	Criar coluna para ordenar os Cursos, por Nível de Formação	DimCurso

Com as transformações anteriores realizadas no ficheiro Excel e no próprio Power BI, os dados encontram-se limpos e preparados para serem carregados. No Power BI, o Editor de Consultas funciona como *data staging area* e os dados já limpos e transformados são carregados no Power BI Desktop, que funciona como *data presentation área*.

4.3.3 Carregamento de dados

Depois de selecionar o botão Fechar e Aplicar, o Editor de Consultas aplica as alterações de consulta feitas, e os dados ficam prontos a usar pelos decisores no Power BI Desktop. Após premido o botão Carregar, é criada uma cópia dos dados a usar nos relatórios [30]. Essa cópia será também enviada para a *cloud* caso se opte por fazer a publicação *online.*

O carregamento no Power BI é completo, ou seja, o botão de atualização elimina os dados e volta a ler da origem. A atualização é manual e feita pelo utilizador, apesar de no Power BI Service ser possível

agendar esta atualização. Neste caso, o carregamento das 10 tabelas de dimensão e 4 tabelas de factos demorou um total de 10 segundos, evidenciando que o impacto no desempenho é reduzido. Existe, contudo, uma possibilidade de visualizar a data da última atualização, bem como de excluir algumas tabelas do carregamento. Com esta opção, as tabelas ficam apenas na *data staging area* e não são carregadas para a *data presentation area*.

Quer no caso dos dados provenientes do CNA, quer no do RAIDES, o carregamento de novos dados ocorre anualmente em setembro e março respetivamente.

Apesar dos dados estarem limpos e transformados, existem melhorias que devem ser feitas para enriquecer o modelo de dados e que, no caso da ferramenta do Power BI, são efetuadas nesta fase, após o carregamento dos dados e concluída a fase de ETL, no Editor de Consultas.

4.4 Enriquecimento do modelo e cálculos analíticos

Uma vez concluídas as transformações no Editor de Consultas, é necessário enriquecer o modelo de dados com outros elementos ou transformações realizadas no Power BI Desktop: gestão das relações, criação de hierarquias, ordenação de dados, agrupamento de dados, obtenção de medidas e colunas calculadas.

4.4.1 Relações

Depois de fechar o Editor de Consultas, o Power BI Desktop cria automaticamente algumas relações entre as tabelas carregadas no modelo. Contudo, o algoritmo utilizado nem sempre deteta todas as relações existentes, pelo que pode ser necessário fazer esta gestão manualmente [19]. A correta criação das relações entre tabelas é um aspeto muito importante uma vez que, qualquer erro cometido aqui, tem a grave consequência de poder vir a gerar informação errada e de estar a falsear os resultados das análises. Esta verificação ou eventual definição de relações em falta, foi feita através do Modo de Relações do Power BI que permite graficamente visualizar a relação entre tabelas ou elementos.

4.4.2 Hierarquias

As hierarquias também devem ser definidas nesta fase de complementar o modelo. As hierarquias em dimensões permitem a análise de medidas em vários níveis de detalhe. A hierarquia relacionada com a dimensão Data, é criada automaticamente pelo Power BI, isto é, desde que no modelo exista um campo de data, o Power BI vai gerar automaticamente diferentes hierarquias de tempo.

Para além da data, foram criadas as hierarquias que tinham sido identificadas na descrição das dimensões (ver Subsecção 4.2.2), e que podem ser visualizadas na Figura 4.7:

Figura 4.7 - Hierarquias criadas nas dimensões Curso, Instituição e País

Para criar uma hierarquia é necessário arrastar e soltar o(s) atributo(s) que se pretende, de forma manual, de modo a ficarem organizados pelos agrupamentos pretendidos.

4.4.3 Ordenação de dados

Para além de gerir as relações e as hierarquias, existem outros procedimentos necessários à melhoria do modelo: um deles tem a ver com a ordenação dos dados de visualização. No caso dos diplomados, este procedimento foi usado para ordenar o atributo Nível de formação por uma ordem lógica e não por ordem alfabética (este procedimento é muito utilizado para ordenar os meses do ano, por ordem temporal e não por ordem alfabética).

Para isso, na dimensão Curso foi criada uma coluna adicional com a ordem pretendida. Esta nova coluna, denominada "Ordem" foi obtida através da criação de uma Coluna Personalizada, ainda na fase de ETL, mas poderia também ser obtida nesta fase de modelação de dados através da opção de criação de uma Nova Coluna. Caso seja na fase de ETL, a coluna aparece no modelo como sendo uma coluna de origem (uma vez que é carregada como tal). Caso seja criada na fase de modelação, fica identificada com o símbolo de uma função.

4.4.4 Agrupamento de dados

Na dimensão Instituição, as Unidades Orgânicas foram agrupadas do modo a distinguir a FCUL das restantes e permitir obter o indicador Percentagem de diplomados com habilitação anterior obtida na FCUL (D3). Assim, o código 0701, que correspondia à FCUL antes da fusão, e 1503, que corresponde ao código oficial e atual da FCUL, foram agrupados como proveniência interna e as restantes faculdades como externa.

Na medida Número de Anos até à Conclusão também foi criado um grupo de modo a agrupar esta variável em apenas duas categorias: N anos e > N anos, sendo N a duração do curso.

4.4.5 Medidas calculadas

Por último, foram criadas as medidas calculadas usando a linguagem DAX. Para criar uma medida, é necessário selecionar a opção Nova Medida no menu Modelação. Automaticamente aparece a barra de fórmulas, em que é possível digitar a expressão DAX que define a medida. Depois de criar uma nova medida, ela aparece numa das tabelas no painel Campos, localizado no lado direito do ecrã, identificada com um ícone de uma calculadora.

Na tabela de factos dos Colocados foi criada a seguinte medida, em linguagem DAX, para dar resposta ao indicador da Média da Nota de Ingresso (A4):

- Média da nota de ingresso = AVERAGE ('FactColocados'[Nota de candidatura])

Na tabela de factos das Inscrições foi criada a seguinte medida, para dar resposta ao indicador Número médio de ECTS inscritos (I3):

- Número médio de ECTS inscritos = AVERAGE ('FactInscrições'[ECTS inscritos])

Por último, na tabela de factos dos Diplomas, foram criadas três medidas, necessárias à obtenção dos indicadores académicos identificados na Subsecção 3.4.4: Classificação Média dos Diplomados (D1), quer a classificação final, quer a parcial e o Número médio de Inscrições até à Conclusão do Curso (D2). As fórmulas utilizadas na sua obtenção foram as seguintes:

- Média da Classificação Final = AVERAGE ('FactDiplomas'[Classificação Final])
- Média da Classificação Parcial = AVERAGE ('FactDiplomas'[Classificação Final MD])

- Média Anos Conclusão = AVERAGE (‘FactDiplomas’[NumInscConclusão])

A vantagem de ter estas medidas, que poderiam ser obtidas apenas em termos de cálculos ao fazer os respetivos gráficos, tem a ver com o desempenho bem como com a possibilidade de voltarem a ser usadas em qualquer visualização. Para isso, basta, como qualquer outra coluna de tabela, arrastá-la e soltá-la no ecrã do relatório ou nos campos de visualização.

4.4.6 Colunas Calculadas

Em termos de colunas calculadas, na tabela de factos dos Diplomas foi criada a Idade de Graduação. Uma coluna calculada é uma nova coluna que transforma ou combina dois ou mais elementos existentes, podendo também servir para estabelecer uma relação entre tabelas. Neste caso, a fórmula, em linguagem DAX, foi a seguinte:

- Idade de Graduação = DATEDIFF(RELATED(DimAluno[Data de Nascimento]);
  FactDiplomas[Data Diploma];YEAR)

Neste exemplo, uma vez que a nova coluna foi criada na tabela de factos dos Diplomas e a data de nascimento consta na tabela Aluno, é necessário utilizar a função *Related* na tabela Aluno.

A idade de uma pessoa é um dado calculado que se determina a partir da data de nascimento. Os dados calculados, de um modo geral, não constam nos sistemas operacionais. Contudo, no *data warehouse* os dados calculados, mesmo os de mais fácil determinação têm sempre lugar, uma vez que simplificam a tarefa do utilizador.

4.5 Relatórios e visualização de dados

A última etapa deste trabalho de BI teve a ver com a visualização e análise da informação, após o carregamento dos dados pretendidos.

Esta secção descreve a fase de criação dos relatórios, bem como a sua publicação no Portal de Ciências.

4.5.1 Desenho e criação dos relatórios

O Power BI Desktop permite criar diversas visualizações de dados, usualmente em relatórios com várias páginas e em que cada página pode ter vários elementos gráficos. Também é possível interagir com as visualizações para ajudar a analisar e compreender os dados, bem como personalizá-las.

Nesta secção pretende-se apresentar a construção do relatório do processo de conclusão, uma vez que os passos utilizados nos relatórios dos outros processos de negócio são semelhantes. Foram seguidas as recomendações de construção de relatórios apresentadas na Subsecção 2.5.2. Pretende-se ainda, com este exemplo, dar resposta às duas perguntas analíticas sobre esta temática, identificadas na Secção 3.2:

- Quais os cursos de licenciatura em que os alunos se graduam com uma média de classificação final mais elevada?
- Qual a média do número de anos até à conclusão dos cursos, por nível de formação?

4.5.1.1 Cursos de licenciatura com uma média de classificação final mais elevada

Para responder a esta pergunta analítica e, uma vez que se pretendia fazer uma comparação, mais especificamente um *ranking*, foi selecionado o gráfico de barras, conforme referido na Figura 2.6 e seguindo a 15ª recomendação sobre a elaboração de relatórios Escolher a visualização apropriada (Figura 2.7) [15].

Em relação à primeira pergunta de negócio, os passos seguidos foram os seguintes:

1) No painel Visualizações escolher o gráfico pretendido, neste caso o gráfico de barras;
2) No painel Campos selecionar, na tabela de factos dos Diplomas, a medida Média Classificação Final e arrastá-la para o campo Valor do gráfico de barras;
3) No painel Campos arrastar o atributo Designação atual do Curso da dimensão Curso, para o campo do Eixo do gráfico de barras;
4) No painel Filtros, no campo Filtros de nível visual arrastar o atributo Descrição do diploma da dimensão Perfil do Diploma e excluir os diplomas parciais;
5) Por último, no painel Visualizações escolher o gráfico denominado segmentações de dados, de modo a criar três segmentações diferentes: nível de formação, ano letivo e departamento responsável. As segmentações de dados são filtros de elementos visuais que permitem a qualquer utilizador segmentar os dados por determinado valor, neste caso por nível de formação, ano letivo e departamento.

O resultado final pode ser visualizado na Figura 4.8:

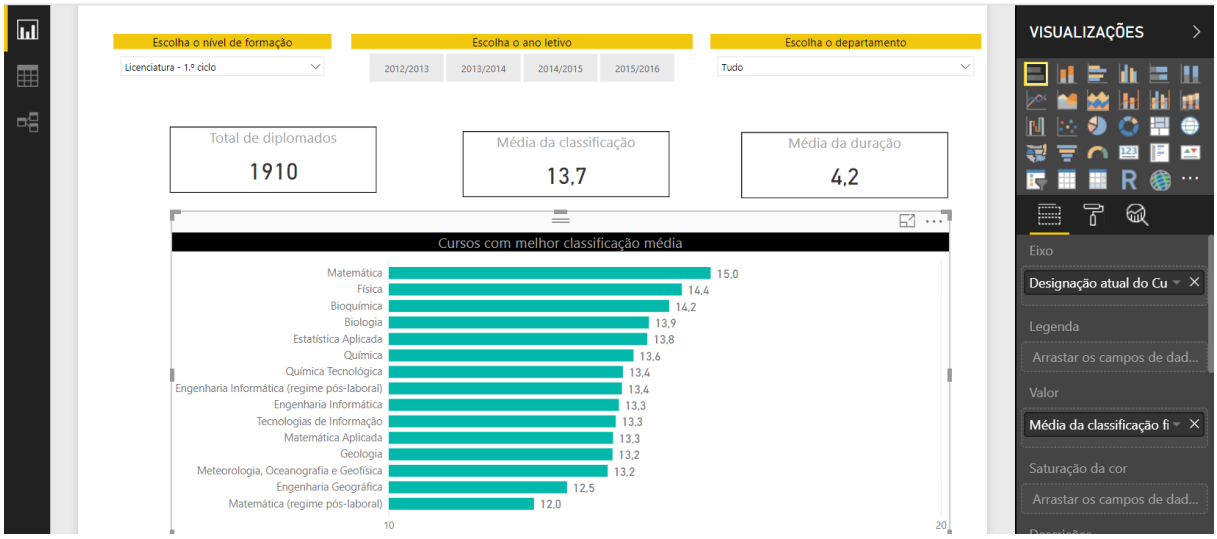


Figura 4.8 - Cursos de licenciatura em que os alunos se graduam com maior média de classificação final

Entre 2012 e 2016, nos cursos de licenciatura graduaram-se 1910 alunos, com uma média de classificações de 13,7 valores, num tempo médio de 4,2 anos. A licenciatura em que os alunos se graduam com uma média final mais elevada é Matemática (15,0), seguida da Física (14,4) e da Bioquímica (14,2).

Para além do gráfico de barras e das segmentações de dados já referidas, utilizou-se outro tipo de visualização denominada cartão de um único número para destacar os dados mais relevantes. No que diz respeito ao conjunto das visualizações, foram consideradas as recomendações sobre a simplicidade, o alinhamento, a escala, os atributos inteligíveis, a uniformização de formatos e de cores.

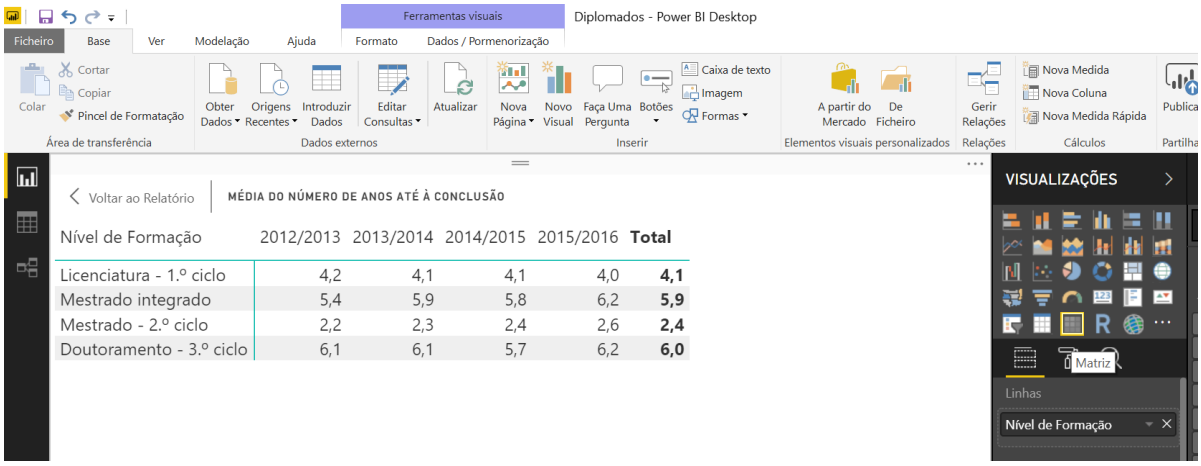
4.5.1.2 Número médio de anos até à conclusão

Para dar resposta à segunda pergunta de negócio, foi necessário:

1. No painel Visualizações escolher o gráfico pretendido, neste caso a matriz;
2. No painel Campos selecionar, na tabela de factos dos Diplomas a medida Média Anos Conclusão e arrastá-la para o campo Valores da matriz;
3. No painel Campos arrastar o atributo da dimensão Curso para o campo Linhas da matriz;
4. No painel Campos arrastar o atributo ano letivo para o campo Colunas da matriz;

5. Por último, no painel Filtros, no campo Filtros de nível visual, no atributo Nível de Formação, excluir as licenciaturas de mestrados integrados, os cursos de especialização e também a licenciatura em Geologia por ter uma duração de quatro anos.

O resultado final pode ser visualizado na Figura 4.9:



Nível de Formação	2012/2013	2013/2014	2014/2015	2015/2016	Total
Licenciatura - 1.º ciclo	4,2	4,1	4,1	4,0	**4,1**
Mestrado integrado	5,4	5,9	5,8	6,2	**5,9**
Mestrado - 2.º ciclo	2,2	2,3	2,4	2,6	**2,4**
Doutoramento - 3.º ciclo	6,1	6,1	5,7	6,2	**6,0**

Figura 4.9 - Número médio de anos até à conclusão, por grau

Entre 2012 e 2016, nos cursos de licenciatura os alunos graduam-se em média em 4,1 anos, mas verifica-se que, em termos evolutivos, este valor tem vindo a diminuir. Nos mestrados integrados este valor situa-se nos 5,9 anos mas, ao contrário das licenciaturas, tem vindo a aumentar. A mesma tendência de aumento é a que se tem vindo a verificar nos mestrados (situando-se a média nos 2,4 anos). Por último, os doutoramentos são o nível de formação onde esta média atinge o valor mais elevado, situando-se nos 6 anos.

4.5.2 Publicação para o *site* de Ciências

Um dos principais requisitos deste trabalho era a publicação dos indicadores no *site* de Ciências, de modo a poder dar resposta às necessidades da comunidade académica.

Uma vez concluídos os diferentes relatórios no Power BI Desktop, e de modo a poder publicá-los na página *web* da Faculdade, foi necessário entrar no Power BI Service, através da opção Publicar.

Após autenticação na conta da Microsoft, através do *email* institucional, e uma vez concluído o carregamento dos dados e do relatório, uma caixa de diálogo informa se o processo de publicação foi bem-sucedido e um *link* é fornecido nessa caixa de diálogo para ser direcionado diretamente para o respetivo relatório no Power BI Service. Caso seja um relatório já existente no Power BI Service, é solicitada a confirmação de substituição do conjunto de dados e dos relatórios anteriores pela versão atualizada.

Em termos de atualização dos dados utilizados neste trabalho as atualizações são pouco frequentes: os dados do Concurso Nacional de Acesso são recebidos da DGES durante o mês de setembro e os do RAIDES em fevereiro/março. Assim, é aceitável o tipo de carregamento do Power BI (apagar os dados antigos e carregar os novos) ainda para mais tendo em conta que o volume de dados utilizado neste trabalho não é demasiado elevado. Existe, contudo, a possibilidade de não Ativar o Carregamento de alguma tabela, caso a mesma não tenha sofrido nenhuma atualização.

No Power BI Service, após ter selecionado o relatório que se pretende publicar, deve escolher-se a opção Ficheiro > Publicar na *Web*, conforme a Figura 4.10:

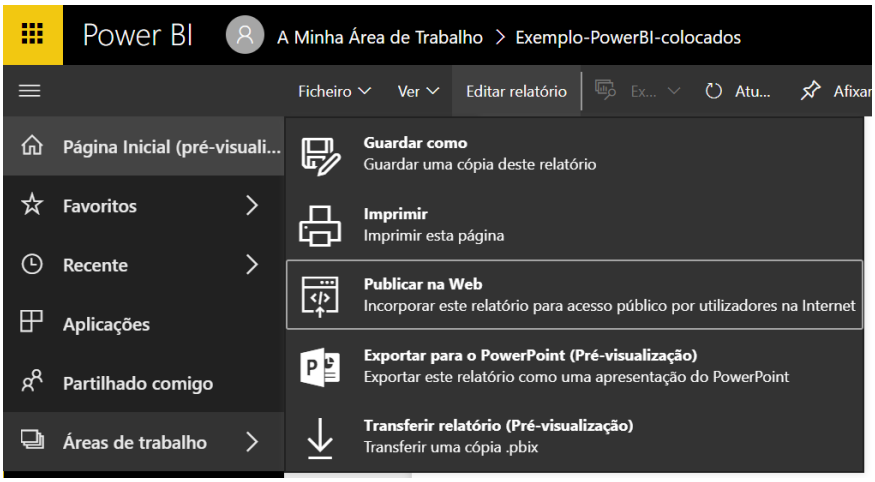


Figura 4.10 - Opção de *Publicar na Web* no Power BI Service

De seguida, aparece uma caixa de diálogo com a informação que será gerado um código de incorporação, que permite publicar o relatório num *site*. Confirmada a opção Criar código de incorporação, o Power BI apresenta uma nova caixa de diálogo, informando que está prestes a publicar um relatório e dados que ficarão públicos, através da *internet*, para qualquer pessoa. Após esta confirmação o Power BI apresenta uma caixa de diálogo com dois *links*:

- Um *link* para partilhar por correio eletrónico, que mostra o relatório como uma página *web*;
- Código HTML (um *link* num *iframe*) que permite inserir o relatório diretamente numa página *web*.

Para o *link* HTML, é possível escolher entre tamanhos predefinidos para o relatório inserido ou modificar o código *iframe* por conta própria e personalizar o seu tamanho.

Seguindo o *link* num navegador, o relatório aparece como uma página *web*, podendo interagir com essa página *web*, exatamente como no Power BI. A Figura 4.11 mostra o exemplo de um dos relatórios, após ter sido inserido o código HTML no Portal de Ciências numa página de Estatísticas - Testes:

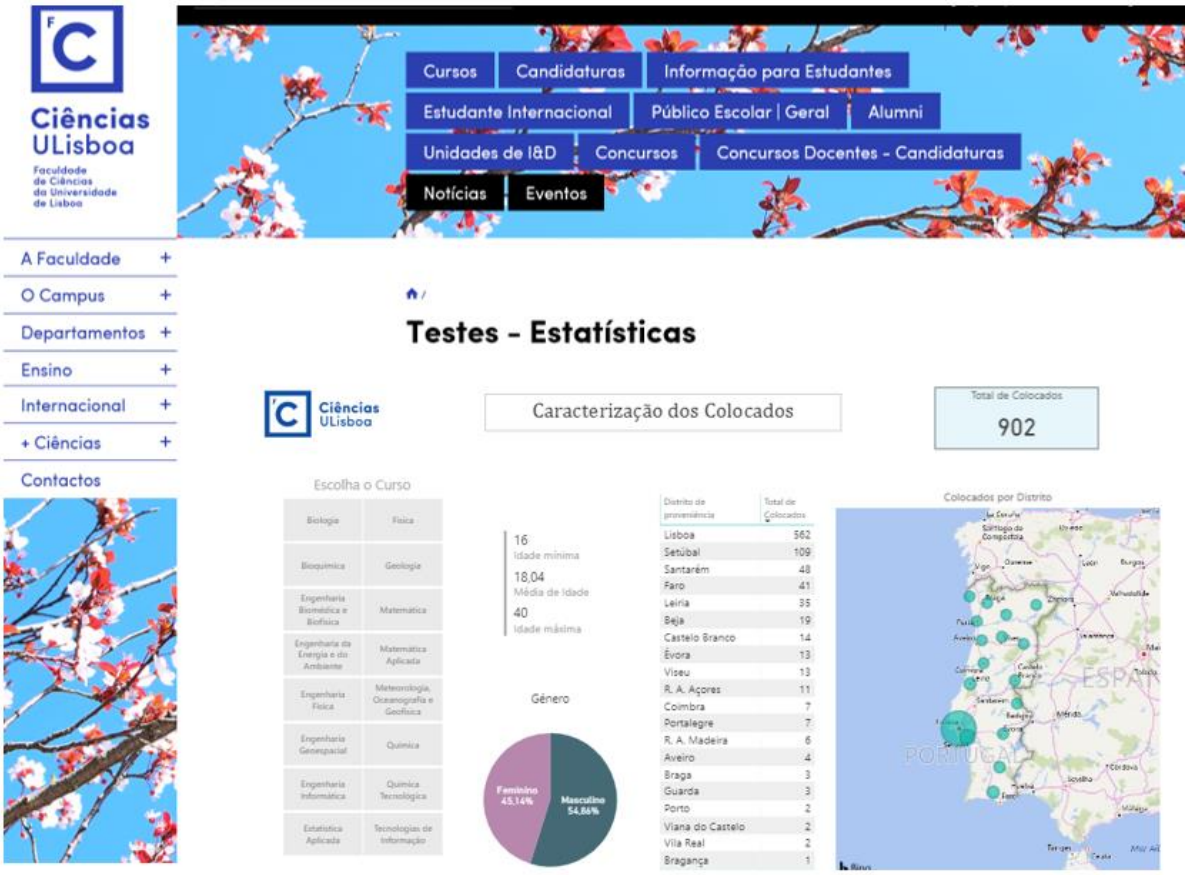


Figura 4.11 - Relatório de colocados no Portal de Ciências

Devido à confidencialidade dos dados dos alunos, a dimensão Aluno não contém nenhum atributo que o possa identificar: nome, número de aluno ou número de identificação. Foram feitos testes que mostram que, através do relatório publicado, os utilizadores não conseguem chegar aos dados em bruto; só conseguem obter os dados que foram utilizados para criar uma determinada visualização. A obtenção dos dados em bruto só é permitida através do Power BI Service, mas, por uma questão de segurança e tendo em conta a aplicação do Regulamento Geral de Proteção de Dados [13], foi tomada esta decisão.

4.6 Avaliação do cumprimento de requisitos

Nesta secção pretende-se fazer uma avaliação do cumprimento dos requisitos identificados na Secção 3.1:

- **Relatórios interativos**: os relatórios finais têm uma visão interativa dos vários atributos, consoante a necessidade de cada utilizador. Existe ainda a possibilidade, no Power BI Desktop, de editar as interações em cada visualização, ou seja, de poder escolher quais as visualizações em que esta interatividade possa não existir. Os *dashboards*, disponíveis apenas no Power BI Service, não possuem esta interatividade.

  De modo a dar resposta aos três tipos de ficheiros que existiam na página das Estatísticas do Portal de Ciências sobre Inscritos, Diplomados e Concurso Nacional de Acesso, foram criados no Power BI três relatórios sobre a mesma temática. Adicionalmente foi obtido um relatório sobre a oferta formativa de Ciências. A Tabela 4.13, apresenta uma descrição mais detalhada do conteúdo de cada um dos relatórios criados:

Tabela 4.13 - Tipos de relatórios criados neste trabalho

Tipo de Relatório	Conteúdo
Candidatos e Colocados	Indicadores académicos sobre caraterização dos alunos e sobre o acesso: indicadores de caraterização dos colocados nos cursos de licenciatura e mestrado integrado, ocupação de vagas, rácios e outras medidas (nota de candidatura, opção).
Inscritos	Indicadores académicos sobre caraterização dos alunos e sobre inscrição: caraterização (nacionalidade, género, idade), número de inscrições dos alunos.
Diplomados	Indicadores académicos sobre caraterização dos alunos e sobre conclusão: caraterização (nacionalidade, género, idade), classificações e número de anos até à conclusão dos diplomados.
Oferta formativa	Indicadores de caraterização dos cursos: por nível de formação, tipologia, ano de avaliação, área CNAEF, área de estudo e duração.

Nos referidos relatórios constam tabelas, gráficos e mapas, para substituir os antigos ficheiros Excel e respetivos gráficos que eram contruídos manualmente em cada ano.

A construção dos diferentes tipos de relatórios deve ser priorizada com base quer nos benefícios esperados (impacto direto, inovação), quer no esforço despendido (existência de dados disponíveis, conhecimento do processo, dimensão do relatório).

- **Existência de filtros:** no Power BI Desktop existem os seguintes tipos de filtros:
  1. Filtros de nível visual: Filtros que podem ser aplicados a cada visualização específica (gráfico, tabela ou mapa);
  2. Filtros de nível de página: Filtros aplicados a todas as visualizações de uma página específica do relatório;
  3. Filtros de nível de relatório: Filtros aplicados a todas as páginas do relatório.

  Adicionalmente aos filtros anteriores existe um tipo de visualização denominado segmentação de dados (*slicer*) que permite a obtenção da informação pelo próprio utilizador, através da escolha de diferentes opções inseridas nos relatórios. Por exemplo:

  - Escolha de um determinado ano letivo, conforme a Figura 4.12, ou visão conjunta de vários anos:

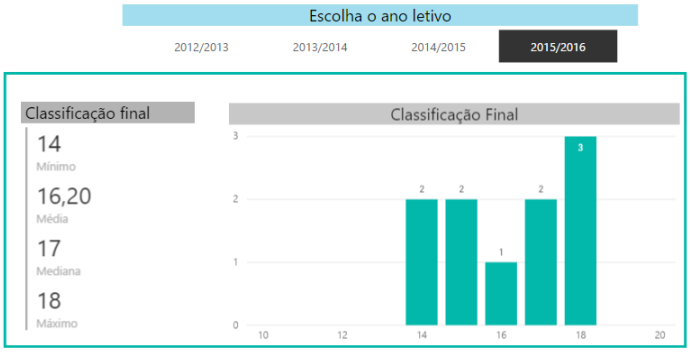


Figura 4.12 - Exemplo de segmentação do ano letivo, através da opção de botões

- Escolha de um determinado nível de formação, curso ou departamento, conforme a Figura 4.13:

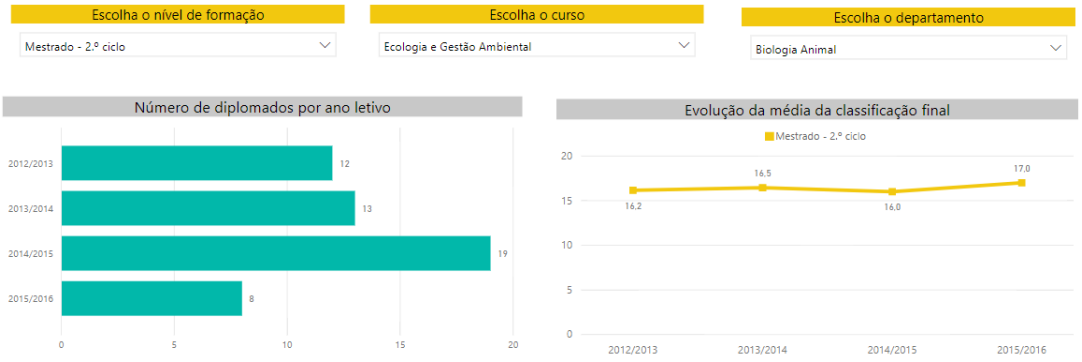


Figura 4.13 - Exemplo de três segmentações de dados, através da opção de lista

Este tipo de visualização permite que as categorias de um atributo sejam apresentadas em forma de botão (exemplo do ano letivo na Figura 4.12) ou em forma de lista (Figura 4.13). Permite ainda que a seleção das categorias seja exclusivamente de um item, de vários itens ou de todos os itens existentes.

- **Existência de hierarquias:** a informação pode ser obtida pelo próprio utilizador com maior ou menor nível de detalhe, através das opções de *drill down* e *roll up (drill up*, na terminologia do Power BI). A Figura 4.14 mostra um exemplo de passar da oferta formativa por grau para a oferta formativa por curso:

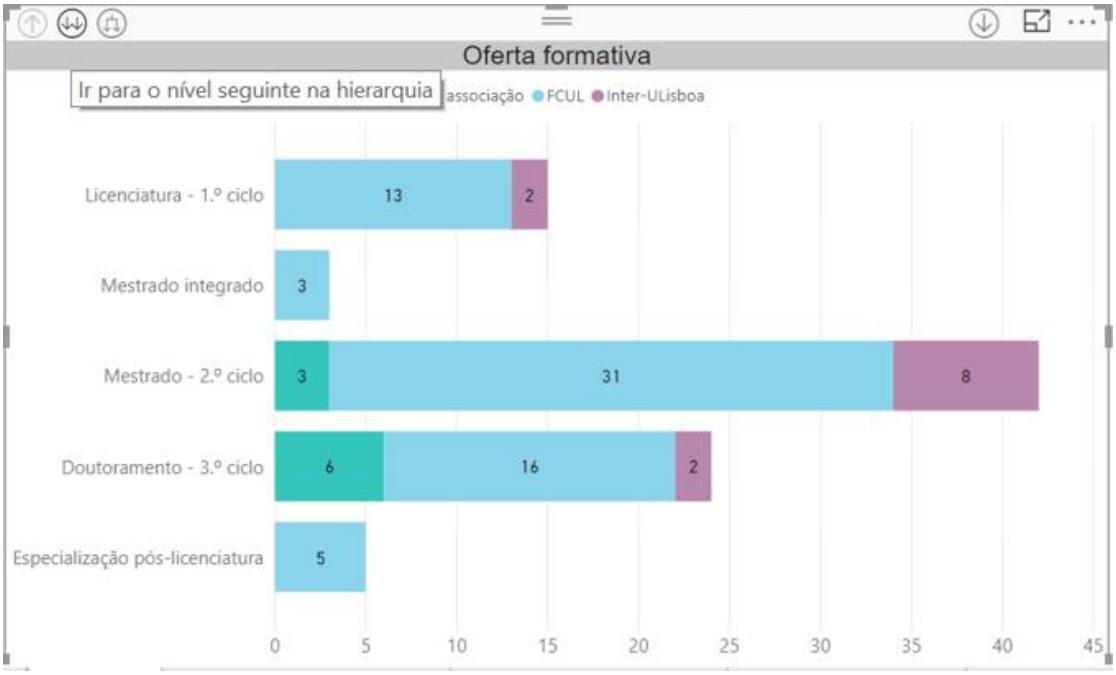


Figura 4.14 - Distribuição da oferta formativa de Ciências, por grau

Ao selecionar o botão Ir para o nível seguinte na hierarquia, e uma vez que tinha sido criada, na Subsecção 4.4.2, a hierarquia Nível de formação > Designação atual do Curso, a visualização passaria a mostrar a oferta formativa por curso, conforme a Figura 4.15. Neste exemplo foi selecionado o nível de formação correspondente à Licenciatura – 1.º ciclo.

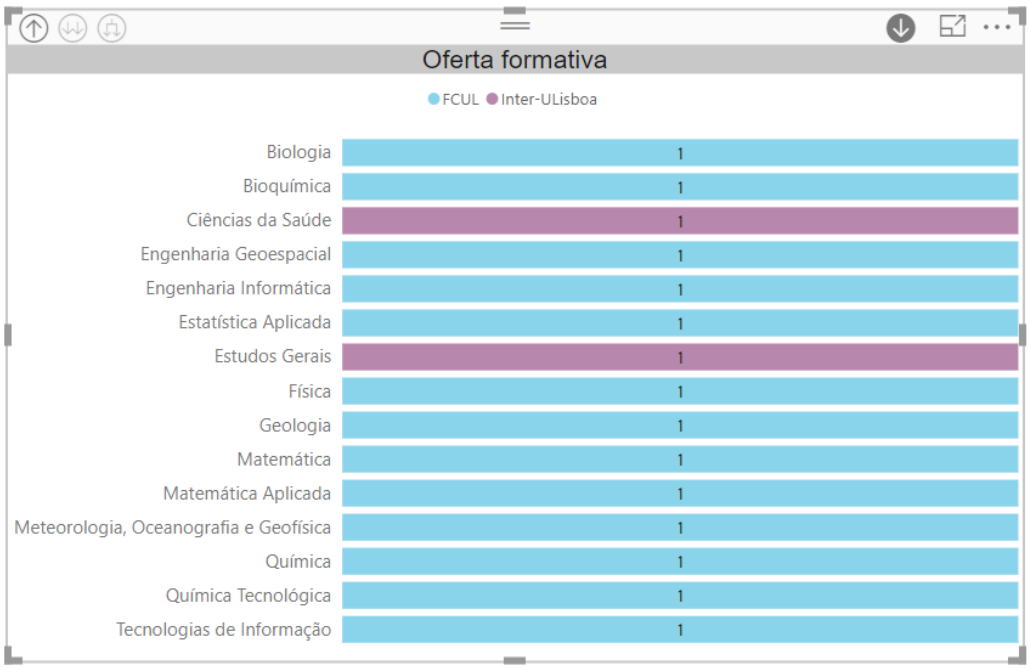


Figura 4.15 - Distribuição da oferta formativa de Ciências por curso de licenciatura

- **Publicação na *web*:** os relatórios podem ser publicados na *web*, conforme mencionado na Secção 4.5.2. Neste caso, os relatórios podem ser disponibilizados no Portal de Ciências, mediante autenticação;
- **Exportação dos relatórios para PDF:** os relatórios podem ser impressos, conforme a Figura 4.10, e consequentemente podem ser exportados para PDF. Existe ainda a possibilidade de Exportar para o PowerPoint. Esta exportação apenas é possível no Power BI Service;
- **Exportação dos dados para ficheiros:** os dados que são utilizados nos diferentes tipos de visualizações (sejam tabelas, gráficos ou mapas) podem ser exportados para um ficheiro de valores separados por vírgulas (CSV), através da opção mais opções > exportar dados disponível no canto superior direito de todas as visualizações, conforme mostra a Figura 4.16:

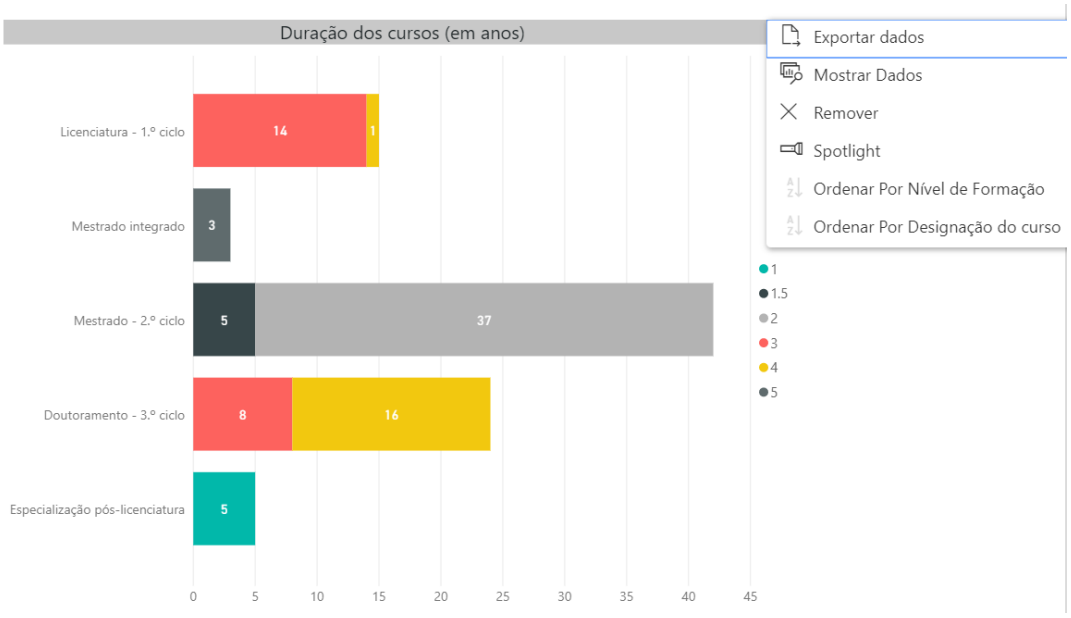


Figura 4.16 - Exemplo da opção de exportação de dados no Power BI Desktop

Esta exportação de dados para ficheiro CSV está disponível no Power BI Desktop. Os relatórios publicados na *web*, têm apenas a opção denominada Mostrar Dados.

Isto significa que todos os pré-requisitos que tinham sido identificados foram cumpridos, apesar de alguns serem possíveis no Power BI Desktop enquanto outros apenas no Power BI Service.

Um último ponto crucial, em termos de requisitos, teve a ver com a avaliação, por parte dos utilizadores, do funcionamento desta plataforma, nomeadamente no que diz respeito à utilização dos relatórios criados. Neste sentido, foi feito um relatório com os dados do Concurso Nacional de Acesso 2018, com o objetivo de avaliar qual o grau de facilidade de obtenção da informação pretendida. A avaliação, feita através de um membro da Direção de Ciências e do Coordenador de um ciclo de estudos, foi bastante positiva. Como pontos fortes foram mencionadas a facilidade de obtenção da informação e a sua disponibilização no próprio dia de divulgação dos dados oficiais pela DGES. Em termos de melhorias, foi feita a sugestão de completar um dos gráficos que não continha explicitamente as percentagens de opção de candidatura, uma vez que não era intuitivo que as referidas percentagens ficavam visíveis ao passar o rato sobre o gráfico.

4.7 Vantagens do novo fluxo de trabalho

No que diz respeito às diferenças em relação ao fluxo de trabalho anterior do GAAI, as dificuldades e os erros ficaram minimizados, conforme se pode avaliar pelos seguintes exemplos que tinham sido identificados na Secção 3.5:

- A dificuldade de obtenção de algumas métricas em Excel (mediana da idade de conclusão por curso e mediana do número de inscrições por curso) ficou muito facilitada no Power BI Desktop, uma vez que agora pode ser conseguida através da escolha da referida medida nas opções da visualização do gráfico, ou como medida calculada através da linguagem DAX. Uma vez obtida, fica de imediato disponível para qualquer curso, através da segmentação de dados por curso.
- Um engano numa determinada fórmula ou num gráfico deixou de implicar alterar ou substituir cada ficheiro e cada gráfico. Com uma correção no gráfico pretendido e uma simples atualização no ficheiro, a nova visualização da informação fica logo disponível. Por exemplo, se num determinado gráfico se pretender excluir o valor de algum dos atributos, basta excluir através dos filtros o referido valor, carregar no botão Atualizar e confirmar que se pretende substituir o ficheiro existente pelo novo;
- A visualização de dados agregados pela designação do curso, em cursos que possuem a mesma designação (como por exemplo a licenciatura em Bioquímica, o mestrado em Bioquímica e o doutoramento em Bioquímica) pode ser realizada através da segmentação de dados por Nível de Formação, sem necessidade de constar o código do curso.
- Passou ainda a existir a possibilidade de ordenar um gráfico por qualquer coluna pretendida, inclusive com colunas que não constem na visualização e personalizadas para este efeito, conforme apresentado na Subsecção 4.4.3.

Assim, com o processo mais facilitado em termos de obtenção de medidas, existe mais tempo para calcular outros indicadores (mais e melhores métricas). Por outro lado, com os processos automatizados, isto é, sem necessidade de serem repetidos anualmente e com a atualização/correção de dados feita com um simples *refresh*, a disponibilização é feita de uma forma muito mais célere. Por último, mesmo com um volume de dados significativo, o desempenho verificou-se excelente, o que significa que ainda é possível adicionar os próximos anos letivos, para continuar a analisar a evolução de determinados indicadores.

4.8 Sumário

Este capítulo apresentou o modelo dimensional da plataforma de indicadores académicos com as tabelas de dimensões e factos; de seguida, descreveu o processo de extração e transformação dos dados, o seu carregamento num repositório único, o enriquecimento do modelo através de determinadas técnicas e o exemplo da elaboração dos relatórios interativos, para o processo da conclusão dos alunos. No fim do capítulo foi feita uma avaliação dos requisitos que tinham sido identificados na Secção 3.1. e uma comparação com o fluxo de trabalho que era realizado anteriormente no Gabinete de Avaliação e Auditoria Interna (GAAI), mencionando as vantagens alcançadas.

5. Conclusões

Neste capítulo são descritas as principais contribuições deste trabalho, as competências adquiridas no decorrer do Projeto, as principais dificuldades encontradas e sugestões para o trabalho futuro.

5.1 Principais contribuições

Os principais benefícios obtidos no desenvolvimento desta plataforma de indicadores académicos, foram os seguintes: melhor apresentação da informação, acesso mais rápido à informação, informação mais concisa, e informação mais específica para diferentes partes interessadas.

Com este trabalho conseguiram-se **automatizar** procedimentos que eram repetidos anualmente e, tendo em conta o contínuo crescimento dos dados, **simplificar** a análise dos mesmos e transformá-los em informação. Esta simplificação e automação permitem ter acesso à informação relevante de uma forma rápida e necessária à tomada de decisão, com poupança de tempo para definição de novas métricas e indicadores. A possibilidade de **analisar tendências** e evoluções era algo essencial para este tipo de dados. Por último, a **disponibilização** em ambiente *web* dessa informação à comunidade de Ciências e de forma interativa em que cada utilizador pode consultá-la de acordo com o seu interesse particular, foi também um contributo muito importante, dado que era um requisito fundamental.

Por último, este trabalho contribuiu para os meus objetivos profissionais e uma vez que estão a ser analisados dados atuais dos alunos da FCUL, os resultados também terão interesse para a comunidade de Ciências, em geral.

5.2 Competências adquiridas

A realização deste trabalho foi uma oportunidade de desenvolver várias competências das quais se destacam as seguintes:

- Desenvolver modelos dimensionais e conhecimentos na área de BI, que é uma área de elevada procura, tendo em conta a necessidade que existe atualmente de trabalhar com grande quantidade de dados. Independentemente da ferramenta de análise, os conhecimentos teóricos desta matéria são essenciais para poder ser aplicados em diversos contextos. Conhecendo os princípios fundamentais de um sistema de BI, foi mais fácil conseguir melhores resultados, independentemente da ferramenta utilizada.
- Aprender a trabalhar nas ferramentas que a Microsoft oferece aos utilizadores de negócios no âmbito da integração, análise e visualização dos dados, quer o Power BI, quer os suplementos de Self Service BI do Excel (Power Query, Power Pivot e Power View). O Power BI Desktop demonstrou ser uma ferramenta flexível, com um bom desempenho e altamente acessível para trabalhar com uma grande quantidade de dados, formatá-los, criar modelos e elaborar relatórios bem-estruturados e mostrou ser a ferramenta adequada para o objetivo pretendido.
- Desenvolver maiores capacidades analíticas, através da linguagem DAX, e de criar diferentes formas de visualização, seguindo as regras de como esta apresentação deve ser feita.
- Aprender boas práticas na escrita, estruturação e formatação de um relatório de Projeto em Gestão de Informação.

Por último, perceber que a melhor solução é uma conjugação de diferentes fatores: por um lado tecnologia, por outro conhecimento do negócio e por último a comunicação com os vários *stakeholders*, para ir de encontro às suas necessidades. Só desta forma é possível ir melhorando resultados.

5.3 Principais dificuldades

Uma das principais dificuldades teve a ver com a limpeza e tratamento dos dados, provenientes das diferentes fontes: a uniformização dos atributos e dos seus valores, a correção/obtenção dos valores em falta, e a decisão sobre o tratamento das mudanças nos atributos de algumas dimensões. Por outro lado, a perceção de quais os atributos que deveriam constar em dimensões ou a necessidade de serem integrados com outros, em minidimensões, também foi um desafio.

A frequente mudança de regras e definições de conceitos do RAIDES também criou alguns obstáculos, inviabilizando algumas comparações anuais e não permitindo a utilização de dados de anos mais antigos, uma vez que não existia comparabilidade em termos de variáveis e campos.

Por outro lado, a dificuldade inicial de dependência da área de Tecnologias de Informação (TI) por causa do repositório de dados ou de conhecimentos de programação, ficou colmatada com o tipo de ferramenta utilizada. Contudo, desta forma os dados ficaram sediados na nuvem e não num servidor da FCUL.

Um sistema de *data warehouse* pode ser complexo, frustrante e difícil de implementar. Os departamentos de TI e as unidades de negócio nem sempre falam a mesma linguagem. A falta de tempo e o número elevado de restrições também se podem converter num obstáculo. Tendo em conta a literatura e a experiência de especialistas desta área, a metodologia que provou ser a melhor opção é fazer pequenas partes para que o utilizador vá vendo o resultado e não esperar pela solução global e completa. Esta foi também a opção considerada neste trabalho: escolher apenas três processos de negócio académicos, de entre uma opção de inúmeros possíveis processos existentes numa Instituição de Ensino Superior.

5.4 Trabalho futuro

Uma vez iniciado este sistema de BI, considera-se que de futuro será muito útil adicionar processos de negócio complementares aos já produzidos neste trabalho e referidos na Secção 3.2. No fundo, o crescimento de um projeto de *data warehouse* deve passar pela inclusão de mais dados e fontes e adaptação a novas necessidades. O sistema de BI desenhado deve ser proativo em relação a necessidades de futuro. Para isso também será necessário que a atual base de dados académica esteja mais estabilizada.

Os processos de negócio para complementar este repositório de dados académicos poderiam passar pela empregabilidade, progresso/avaliação (abandono/sucesso escolar) e satisfação (inquéritos pedagógicos). Com os indicadores destes processos já seria possível a obtenção de um relatório de curso, necessário à certificação do Sistema Interno de Garantia da Qualidade da FCUL.

Numa visão mais a longo prazo, a comparação dos indicadores do acesso com outras Unidades Orgânicas e Instituições de Ensino Superior também seria relevante.

Os pontos anteriores têm a ver com o crescimento do sistema de BI. Por outro lado, a manutenção do sistema e do repositório dos dados também é fundamental, dado que o projeto não fica concluído com a solução inicial. A contínua exigência ao nível da qualidade dos dados é também outro ponto crucial para a credibilidade do sistema no presente e no futuro.

Bibliografia

[1] Andreas Schleicher, *Todas as escolas vão poder definir um quarto do currículo*, Semanário Expresso de 30 de abril de 2016 - Primeiro Caderno, p.21

[2] José L. Pereira, *Tecnologia de Bases de Dados*, FCA, 1997

[3] Lei n.º 38/2007 de 16 de agosto, Diário da República – n. º157/07 - I Série, Assembleia da República

[4] Vidal de Carvalho, Ana Azevedo, António Abreu, *Bases de dados com Microsoft Access 2007*, Centro Atlântico, 2008

[5] Salvador Ramos, *Business Intelligence (BI) & Analytics*, SolidQ, 2016

[6] Hugh Watson, George Houdeshel, e Rex Rainer, *Building executive information systems and other decision support applications*, Wiley, 1997

[7] António Ferreira, *Integração e Processamento Analítico de Informação*, guião das aulas teóricas, 2017

[8] Salvador Ramos, *Microsoft Business Intelligence: vea el cubo medio lleno*, SolidQ, 2011

[9] William H. Inmon, *Building the data warehouse*, 4ª edição, Wiley, 2005

[10] Ralph Kimball e Margy Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*, Wiley, 2002

[11] Cole N. Knaflic, *Storytelling with data: a data visualization guide for business professionals*, Wiley, 2015

[12] Naciones Unidas, *Cómo hacer comprensibles los datos – Parte 2 – Un guía para presentar estadísticas*, Comisión Económica para Europa, 2009

[13] Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, 27 de abril de 2016, [em linha]. Disponível em https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=PT [Acedido em julho de 2018]

[14] Naciones Unidas, *Cómo hacer comprensibles los datos – Parte 1 – Un guía para escribir sobre números*, Comisión Económica para Europa, 2009

[15] Marco Russo, *Power BI visualization best practices* [Webinar de 31/05/17] [em linha]. Disponível em https://community.powerbi.com/t5/Webinars-and-Video-Gallery/5-31-2017-Power-BI-visualization-best-practices-by-Marco-Russo/m-p/161482?Is=Website

[16] Cindi Howson, Rita L. Sallam, James Richardson, João Tapadinhas, Carlie J. Idoine, Alys Woodward, *Magic Quadrant for Business Intelligence and Analytics Platforms*, Gartner, 2018 [em linha]. Disponível em https://www.gartner.com/doc/3861464/magic-quadrant-analytics-business-intelligence [Acedido em maio de 2018]

[17] Stratebi, *Análises de herramientas BI* [em linha]. Disponível em http://www.todobi.com/2017/04/comparativa-de-herramientas-business.html [Acedido em abril de 2017]

[18] Microsoft, *Aprendizagem orientada – Power BI* [em linha]. Disponível em: https://powerbi.microsoft.com/pt-br/guided-learning/ [Acedido em setembro de 2018]

[19] Alberto Ferrari, Marco Russo, *Introducing Microsoft Power BI*, Microsoft Press, 2016

[20] Lei n.º 62/2007 de 10 de setembro, *Diário da República n. º174, I série*, Assembleia da República

[21] Direção-Geral do Ensino Superior, *Concurso nacional de acesso – 2017 em números* [em linha]. Disponível em http://www.dges.gov.pt/estatisticasacesso/2017/ [Acedido em março de 2018]

[22] Direção-Geral de Estatísticas da Educação e Ciência, *Documento técnico da Plataforma de Recolha de Informação do Ensino Superior – RAIDES*, [em linha]. Disponível em: http://www.dgeec.mec.pt/np4/raides17/%7B$clientServletPath%7D/?newsId=879&fileName=RAIDESPRIES.pdf [Acedido em fevereiro de 2018]

[23] Direção-Geral de Estatísticas da Educação e Ciência, *Inquéritos > RAIDES*, Disponível em: http://www.dgeec.mec.pt/np4/raides [Acedido em janeiro de 2018]

[24] Portaria n.º 181-D/2015 de 19 de junho, *Diário da República n. º118, I série*, Ministério da Educação e Ciência

[25] A3ES, *Referenciais para os sistemas internos de garantia da qualidade nas instituições de ensino superior*, 2016. Disponível em: http://www.a3es.pt/pt/node/77182 [Acedido em outubro de 2017]

[26] Cláudia S. Sarrico, *Indicadores de desempenho para apoiar os processos de avaliação e acreditação de cursos*, Gabinete de Estudos e Análise - A3ES, 2010

[27] Salvador Ramos, *Excel 2013, Power Pivot y DAX*, SolidQ, 2016

[28] Maribel Yasmina Santos e Isabel Ramos, *Business Intelligence – Da informação ao conhecimento*, FCA, 2017

[29] Ricardo Garcia, Maria João Valente Rosa e Luísa Barbosa, *Que número é este? Um guia sobre estatísticas para jornalistas*, Fundação Francisco Manuel dos Santos - Pordata, 2017. Disponível em: https://www.pordata.pt/ebooks/GuiaJornalistas/mobile/index.html#p=1 [Acedido em junho de 2018]

[30] Alberto Magalhães, *Business Intelligence no SQL Server*, FCA, 2017